Classifying Semantic Clause Types With Recurrent Neural Networks: Analysis of Attention, Context & Genre Characteristics

Maria Becker* — **Michael Staniek*** — **Vivi Nastase*** — **Alexis Palmer**** — **Anette Frank***

* *Heidelberg University, Department of Computational Linguistics*
** *University of North Texas, Department of Linguistics*

ABSTRACT. *Detecting aspectual properties of clauses in the form of semantic clause 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 linguistic features. We introduce an attention mechanism that pinpoints relevant context information. Our model implicitly captures task-relevant features and avoids the need to reproduce explicit linguistic features for other languages.We present experiments for English and German that achieve competitive performance, and analyze the outputs of our systems from a linguistic point of view. We present a novel take on modeling and exploiting genre information and showcase the adaptation of our system from one language to another.*

RÉSUMÉ. *Il a été démontré que la détection des propriétés aspectuelles des clauses sous la forme de clauses sémantiques dépend d'une combinaison de caractéristiques linguistiques. Nous explorons cette tâche dans un cadre d'apprentissage sur la base des réseaux de neurones profonds. Nous introduisons un mécanisme d'attention qui identifie le contexte pertinent. Notre modèle permet d'éviter la nécessité de reproduire des caractéristiques linguistiques pour d'autres langues. Nous présentons des expériences pour l'anglais et l'allemand qui atteignent des performances compétitives, et explorons nos résultats d'un point de vue linguistique. Nous présentons une nouvelle approche pour la modélisation de l'information du genre de texte et nous mettons en valeur l'adaptation de notre système d'une langue à l'autre.*

KEYWORDS: *semantic clause types, situation entities, deep-learning, GRU, embeddings, attention mechanism, sequential information, genre, English, German.*

MOTS-CLÉS : *types de clauses sémantiques, entités de situation, deep-learning, URG, word embedding, mécanisme d'attention, information séquentielle, le genre de texte, anglais, allemand.*

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). Sophisticated features have been built by the prior work to capture diverse linguistic indicators of SE-types, including morpho-syntactic and rich semantic features. The larger context was also shown to be useful: Friedrich *et al.* (2016) used a sequence labeling approach that took into account contextual clause labels, leading to improved classification performance.

Deep learning provides a powerful framework in which linguistic and semantic regularities can be implicitly captured (to a certain degree) through word embeddings (Mikolov *et al.*, 2013b). Also, neural systems are able to detect features useful for a given task while learning takes place, through error back-propagation (Goodfellow *et al.*, 2016; Goldberg, 2017). 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., SE-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) a setting approximating sequential labeling in which the attention vector provides features that describe the clauses preceding the target 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, context representation using labels of previous clauses, and information about the text genre significantly 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. Since our system is supervised, this presupposes that annotated training data is available.

A downside of neural models is of course that they are relatively opaque with respect to the nature of the learned features. We try to counter this weakness by deeper model analysis: by exploiting the weights of the learned attention vectors and by

relating genre-specific linguistic properties to the classification performance of genre-aware SE classification models.

Our work presents a novel take on modeling and exploiting genre information for the task of SE classification. We test our models on the English multi-genre corpus of Friedrich *et al.* (2016) and on a German multi-genre corpus which we assembled for that purpose. We provide qualitative evaluation and investigation of the learned models, by relating learned attention weights of the model to different linguistic attributes: POS classes, individual word tokens and positional information. We also investigate genre-specific information, such as frequent SE-type n-grams in different genres, and relate them to the performance of the genre-aware SE classification models.

Our aims and contributions are: (i) We study the performance of GRU-based models enhanced with attention over various window sizes 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 model extensions that integrate external knowledge about genre and show that this improves 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 and show that the performance is comparable to English. (v) We perform qualitative evaluation based on learned attention weights and distributional information on genre.

In what follows, Section 2 introduces the linguistic categories of semantic clause types as used in our work. Section 3 situates our contribution in relation to prior work on linguistic and computational aspects of SE-type classification. Section 4 proposes GRU-based model variants for SE-type classification, including local and context-informed models, models that incorporate attention over token embeddings or predicted SE-type labels in the previous context, and models that make use of external genre information as an additional feature. In Section 5 and 6, we describe our experimental data and settings, how we train and evaluate our models, and report and compare results. Section 7 presents a deeper investigation of the learned attention weights and the impact of textual genre for SE classification. In Section 8 we transfer our system to annotated data for German and analyze its performance, also in relation to the English classifier. Section 9 summarizes and concludes with perspectives on future work.

2. Semantic Clause Types

Situation entities were identified by Smith (2003) as one of the linguistic correlates to variations in text type at the level of the text passage. In other words, modes of discourse such as *narrative* and *argument/commentary* are distinguishable from one another by readers in part because of their varying distributions of situation types (or semantic clause types). Narrative passages consist primarily of events and states, for example, and argumentative passages make heavy use of generics and generalizing

sentences. These observations have since been supported by further empirical investigations (Becker *et al.*, 2016a; Mavridou *et al.*, 2015; Palmer and Friedrich, 2014).

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 Palmer *et al.* (2007) and Friedrich and Palmer (2014b). SE-types describe the abstract semantic types of situations evoked in discourse through clauses of text. As such, they capture the manner of presentation of the content, along with the information content itself. For example, some propositional content can be alternately described with a focus on the eventive aspect of the proposition (*The car squealed around the corner*) or on the stative aspects of the proposition (*The car's squeal was deafeningly loud*).

The seven SE-types we use are described below. The first subset – eventualities – consists of **states**, **events**, and **reports**. Report-type entities (e.g., the italicized portion of (3) below) typically provide attribution for statements and are modeled as a subtype of event.

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.

Further SE categories are **generic sentences** and **generalizing sentences** (sometimes referred to as habituals). The former predicate over classes or kinds; the latter describe regularly-occurring events, such as habits of individuals.

4) GENERIC SENTENCE (GEN): *Birds can fly. – Scientists make arguments.*
5) GENERALIZING SENTENCE (GS): *Fei travels to India every year.*

The final two SE-types included in our inventory are QUESTION and IMPERATIVE.

6) QUESTION (Q): *Why do you torment me so?*
7) IMPERATIVE (IMP): *Listen to this!*

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

Semantic features for SE-type. Determining the SE-type is a complex task involving interactions between lexical and grammatical information, syntactic structure, and various aspectual and other semantic features. In particular, three classes of semantic feature have been identified as useful for identifying the SE-type of a clause (Friedrich and Palmer, 2014b): the *lexical aspectual class (stative or dynamic) of the clause's main verb*, *habituality of the clause*, and the *nature of the main referent of the clause*. [1] An especially useful main referent feature is the *genericity of the main referent* – whether or not it evokes a class or kind.

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

Motivation. The semantic distinctions made by this inventory of SE-types are linguistic in nature, and each individual SE category has been well-studied in linguistic theory; please see Smith (2005) for an extensive list of relevant literature. The particular inventory is motivated by theoretical work which aims to understand the nature of text type. Smith (2003) assembles this set of clause types following investigation of the linguistic differences between text passages of different Discourse Modes (e.g., argumentative, narrative, reporting), because these clause types, together with mode of progression, explain distinctions between text types. In addition, they allow for near-exhaustive annotation of the clauses of an individual text passage. In this work, we use the inventory described on the previous page, which is a subset of Smith's inventory, leaving out only two infrequently-occurring types in the category of ABSTRACT ENTITIES (Friedrich *et al.*, 2016).

The ability to classify clauses by SE +-type lays the foundation for automatic classification of text passages according to Discourse Mode (see, for example, Song *et al.* (2017)). In addition to their role in text type classification, SE-types have been shown to be useful for determination of event duration (Vempala *et al.*, 2018; Sanagavarapu *et al.*, 2017). Additional applications are anticipated in temporal interpretation, event extraction, and narrative analysis as well as for the extraction of knowledge, e.g., generalizing knowledge (Reiter and Frank, 2010). Moreover, SE-types play a role in argumentation structure analysis (Becker *et al.*, 2016b) and have been shown to be useful for genre characterization (Palmer and Friedrich, 2014).

3. Related Work

Semantic clause types and text passages. The use of linguistic features for distinguishing text passages is closely related to Argumentative Zoning (Teufel, 2000; O'Seaghdha and Teufel, 2014), where linguistic features are used to distinguish genre-specific types of text passages in scientific texts. In this manner, those texts are segmented into types of text passages such as Methods or Results. 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). Notions related to SE-types have been widely studied in theoretical linguistics (Vendler, 1957; Verkuyl, 1972; Dowty, 1979; Smith, 1991; Asher, 1993; Carlson and Pelletier, 1995) and have seen growing interest in computational linguistics (Siegel and McKeown, 2000; Zarcone and Lenci, 2008; Herbelot and Copestake, 2009; Reiter and Frank, 2010; Costa and Branco, 2012; Nedoluzhko, 2013; Friedrich and Palmer, 2014a; Friedrich and Pinkal, 2015; Song *et al.*, 2017).

Feature-based classification of SE-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). Syntactic features include (among others) selected structural configurations associated with SE-type, as well as dependency relations associated with the main referent. Semantic features include (among others) WordNet senses, countability, and

presence of negation and/or modality. 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 (cf. 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, noting 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 to 69.8. This shows that their classifiers generalize over the different genres present in the dataset. However, genre information is not explicitly modeled in their approach.

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.*, 2014; Mishra *et al.*, 2017) which are particularly suitable for tasks that rely on discovering patterns that are distributed over the input signal. RNN variations – with Gated Recurrent Units (GRU) (Cho *et al.*, 2014; Abdul-Mageed and Ungar, 2017) 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)). Sentence representation learning from specifically selected training data has also been done using bi-LSTM models (Conneau *et al.*, 2017; Nie *et al.*, 2017). Tree-based LSTM models have been shown to often perform better than purely sequential bi-LSTMs (Tai *et al.*, 2015; Miwa and Bansal, 2016; Cheng and Miyao, 2017), but depend on parsed input.

Hierarchical classification models. Song *et al.* (2017) develop a neural hierarchical multi-class sequence labeling model for automatic labeling of Discourse Modes in essays. Their system uses a sentence-level GRU layer for sentence encoding and a bi-

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

GRU layer to connect the encoded sentences and to perform sequence prediction for discourse mode labeling. They use this model to improve automatic essay scoring in Chinese using the predicted discourse mode labels as features. Their work is related to, but by-passes, the level of SE classification. Their model does not make use of the attention mechanism and does not offer a deeper linguistic analysis of the learned models.
Song *et al.* (2017) observe that accessing information about past and future sentences provides more contextual information for current discourse mode prediction, which is in line in with our hypothesis that modeling contextual information yields improved performance for classifying semantic clause types. The model we propose in Section 4 incorporates context information by using separate GRUs and predicts the SE-type for one clause each time. Inspired by Song *et al.* (2017)'s work, we leave a model which jointly learns representations for sequences of clauses in a text or a paragraph as future work. [3]

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 make use of attention to larger context windows and previous labeling decisions to capture sequential information relevant for our classification task, and we investigate the learned weights to gain insights about what the models learn.

4. Models

We aim for a system that can fine-tune input word embeddings to the task, and that 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 process successfully 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.

3. During the revision phase of this article, Dai and Huang (2018) published a hierarchical neural model for SE classification. They design a unified neural network which models word-level dependencies and clause-level dependencies jointly in order to derive clause representations for SE-type prediction. When being trained on the English dataset which we also use in our work, this model achieves up to 80.7 accuracy on the test set, beating all of the baseline models. We expect that adopting a hierarchical classification framework will result in further improvement of our results.

4.1. *Model Components*

4.1.1. *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. The latter makes the representations suitable for our task, given that we aim to capture sequence information. 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) compared to 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. Comparison of GRUs, bi-GRUs, LSTMs and bi-LSTMs on our dataset for our classification task – in initial experiments, not reported here – showed that GRUs outperform the other three, confirming this hypothesis.

The relevant equations for a GRU are given below. x_t is the input at time t (usually a dense word embedding vector), 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.

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad [1]$$

$$\tilde{h_t} = tanh(W x_t + U(r_t \odot h_{t-1})) \quad [2]$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad [3]$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h_t} \quad [4]$$

The last hidden vector h_T (with T the number of tokens in the input clause) 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 $softmax$ to transform it to a probability distribution.

4.1.2. *Neural Attention Mechanism*

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 base implementation of attention used in Rocktäschel *et al.* (2016) for our clause classification task as follows:

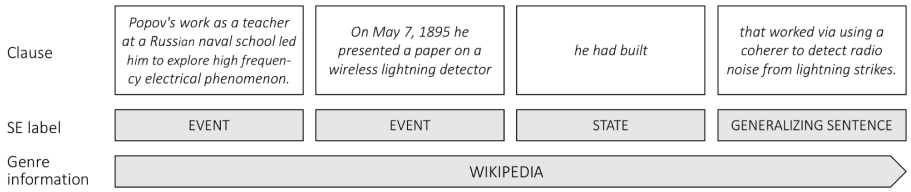

Figure 1. *An example from Wikipedia illustrating context and genre information modeled in our system. Assuming "he had built" is the clause to be classified (with label* STATE*), context and genre information can be taken into account in different ways: the* ***token context model*** **CON_TOK2+GEN** *uses all tokens of the previous two clauses jointly with information about the genre (here, Wikipedia); the* ***label context model*** **CON_LAB2+GEN** *instead uses as input the target clause and the (predicted) labels of the two previous clauses (*EVENT, EVENT*) jointly with genre information.* ***Local models*** *would use only the target clause token inputs (optionally jointly with genre information) for classification.*

$$M = tanh(W_h H + W_v h_T \otimes e_T) \quad [5]$$

$$\alpha = softmax(w^T M) \quad [6]$$

$$r = H\alpha^T \quad [7]$$

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_T is a vector of 1s where T denotes the T tokens in the input clause. $\otimes$ denotes the outer product of the 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 h_T.

$$h^* = tanh(W_p r + W_x h_T) \quad [8]$$

where W_p and W_x are trained projection matrices. We convert h^* to a real-valued vector of length 8 (the number of target classes) and apply $softmax$ in order to transform it to a probability distribution over the 8 output classes, i.e., the predicted SE-types.

4.1.3. *Modeling Context and Genre Information*

Previous analyses show that text types differ with respect to their SE-type distributions (Friedrich and Pinkal, 2015). Furthermore, specific n-grams over SE-types are

more frequent within some textual genres than in others. This supports the choice of incorporating (sequential) context information and information about genre as additional features for the classification of SE-types. The English corpus we use consists of texts from 13 genres; the German corpus covers 7 genres. Figure 1 illustrates both the context and the genre information that our models consider for classifying SE-types.

4.2. *Model Types*

We investigate different model types: Local Models and Context Models.

4.2.1. *Local Models*

LOC, LOC_ATT and LOC_ATT+GEN. We first experiment with models that only consider the local clause for SE classification. Adding attention mechanism and genre information to our basic local model results in three versions of the local model: (i) **basic local model** (LOC) uses as input only the hidden representation computed over the tokens of the clause to be classified; the last hidden vector $[h_T]$ (h_T from Equation 4) is passed through the fully connected layer and softmax; (ii) **local model enhanced with attention** to the representations of the tokens in the local clause (LOC_ATT); here $[h^*]$ (h^* as defined in Equation 8) is used for projection; and (iii) **local model enhanced with attention and genre information** (LOC_ATT+GEN), where genre information is encoded as a dense embedding g of a genre label which is initialized randomly [4]; here we concatenate the attention vector h^* and a genre label embedding g $[h^*; g]$. Illustrations for all three model types are given in Figure 2.

4.2.2. *Context Models*

We also investigate models that consider not only the local clause for SE classification in model training, but also the previous clauses' token sequences or their labels. We experiment with several settings:

1) different window sizes of token sequences or of previous labels

2) applying or omitting attention to token sequences or previous labels

3) adding or omitting genre information.

These settings result in various model combinations. In the interest of clarity and space, we only describe and report the results for our best performing models for the following three categories: (i) context models using *tokens* of previous clauses jointly with genre information; (ii) context models using *labels* of previous clauses jointly with genre information; and (iii) context models using *tokens and labels* of previous clauses jointly with genre information.

4. We also use a GRU for the genre representation. This was a design decision in order to keep representations uniform. Note that the individual GRUs (tokens, labels, genre) do not share parameters.

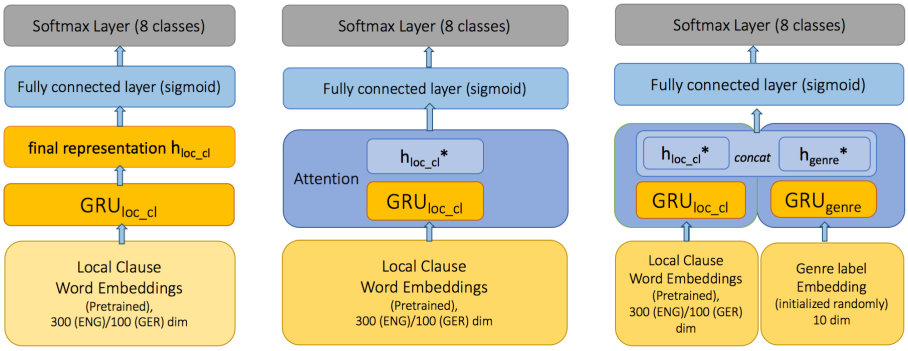

Figure 2. *Architecture of our local models: basic local model without attention (*LOC*, left), local model with attention (*LOC_ATT*, middle) and local model with attention using genre information (*LOC_ATT+GEN*, right).*

CON_TOK+GEN. When considering **tokens of previous clauses**, we add one GRU model for each previous clause ($h_1; h_2; ...; h_N$, with N the number of previous clauses) and concatenate their final outputs with the final output of the GRU with attention for the target clause h_0^* and with the final output of the GRU for genre label encoding h_g (cf. Figure 3).[5]

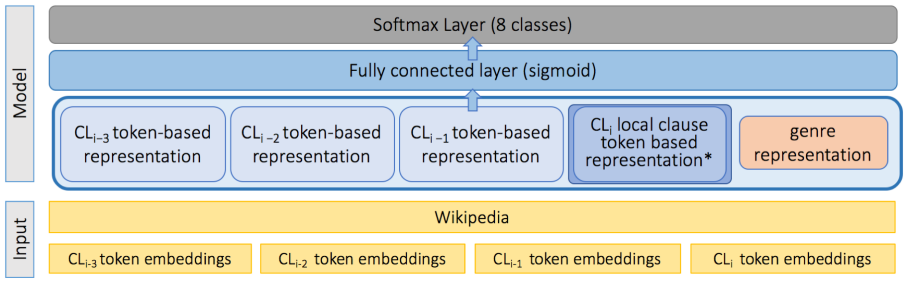

Figure 3. *Context model using tokens of target clause (with attention), tokens of previous clauses (without attention) with N=3, jointly with genre information (*CON_TOK+GEN*).*

In our experiments we found that models perform best when we apply the attention mechanism only to the GRU for the target clause itself (instead of applying it also to the GRUs for the previous clauses).

$$h^*_{con\_tok+gen} = [h_1; h_2; ...; h_N; h_0^*; h_g] \quad [9]$$

5. The concatenation operation is denoted by square brackets, and the elements which are concatenated are separated by semicolons.

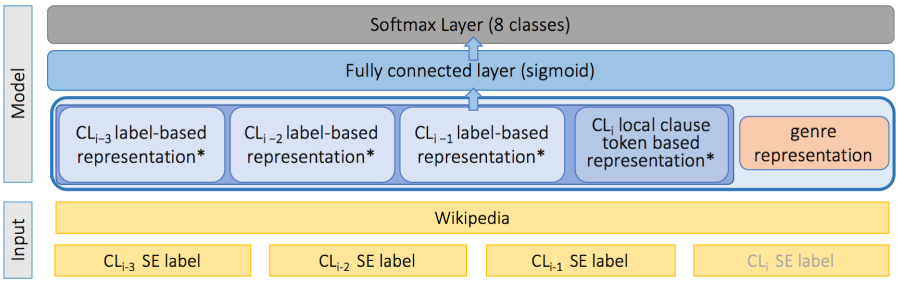

Figure 4. *Context model using tokens of target clause (with attention), labels of previous clauses (with attention) with N=3, jointly with genre information (*CON_LAB+GEN*).*

We then transform the concatenated vector into a dense vector equal to the number of class labels and apply $softmax$ to predict the local clause's SE-label.

CON_LAB+GEN. For including **labels of the previous clauses** in our model, we first transform the gold-standard labels used during training into embeddings, concatenate them and apply attention to the sequence of labels (i.e., to the concatenation of the label vectors). We then concatenate the final hidden state of the target clause with the attention vector learned over the sequence of labels of the previous clauses and with the final output of the GRU for genre, cf. Figure 4:

$$h^*_{con\_lab+gen} = [h^*_{lab}; h^*_0; h_g] \quad [10]$$

where h^*_{lab} is the last hidden state from the GRU used on the label sequence of previous labels over which attention is applied jointly ($h^*_{lab} = [h_1; h_2; ...; h_N]^*$), h^*_0 is the final output of the GRU with attention for the target clause, and h_g is the final output of the GRU for genre. At test time, we use the predicted probability distribution vector of the labels of the previous clauses.

$$h^*_{con\_toklab+gen} = [h_1; h_2; ...; h_N; h_{lab}; h_0; h_g] \quad [11]$$

CON_TOKLAB+GEN. We also perform experiments that include both the embedding representations for tokens and for the labels of previous clauses. One GRU model is added for each of the previous clauses (tokens), their final outputs $h_1; h_2; ...; h_N$ are then concatenated with the embeddings for the labels of the previous clauses h_{lab}, with the final output of the GRU for the target clause h_0, and with the final output of the GRU for genre h_g. This model performs best when the attention mechanism is omitted. It is illustrated in Figure 5.

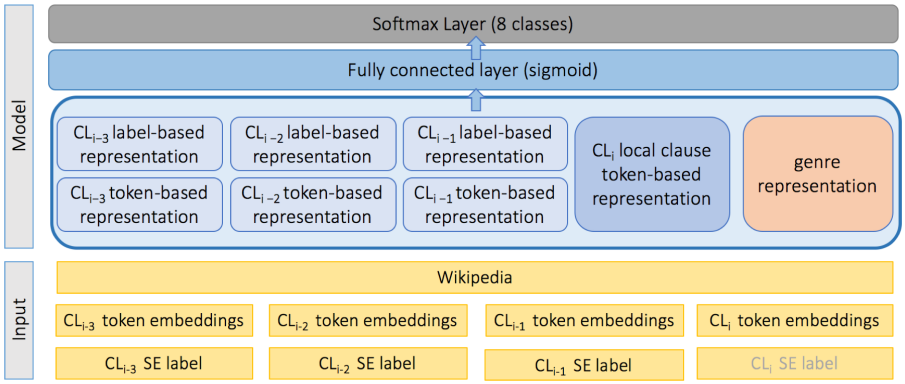

Figure 5. *Context Model using tokens of target clause (without attention), tokens and labels of previous clauses (both without attention) with N=3, jointly with genre information (*CON_TOKLAB+GEN*).*

5. Data

5.1. *Datasets*

We use the English dataset described in Friedrich and Palmer (2014b).[6] The texts, obtained 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 three data sets described in Mavridou *et al.* (2015), Becker *et al.* (2016a) and Becker *et al.* (2017).[7] The German texts cover 7 genres: argumentative essays (Peldszus and Stede, 2015), Wikipedia articles, fiction, commentary, news texts, TED talks, and economic reports. Statistics are given in Table 1.

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

Table 1. *Datasets with SE-labeled clauses*

Figure 6 gives an overview of the distribution of instances (i.e., clauses) among genres within our English and German datasets. Compared to the English dataset, the German dataset is smaller (44% in size) and less diverse with respect to genre

6. Available at: https://github.com/annefried/sitent.

7. Available at: http://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/GER_SET/GER_SET_data.shtml.

(7 instead of 13 genres). The genres in the German dataset are more similar to one another than those in the English dataset.

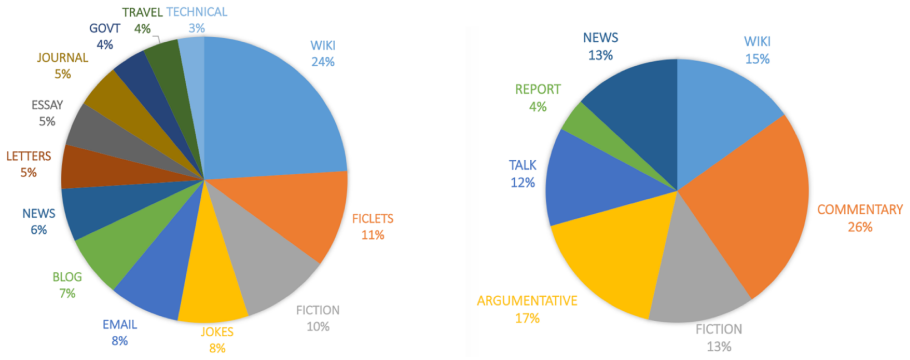

Figure 6. *Distribution of instances among genres within our English (left) and German (right) datasets.*

5.2. *Distribution of SE Types and their N-grams among Different Genres*

Text types differ in their SE-type distributions: Palmer and Friedrich (2014) 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.

The distribution of SE-types in our datasets. When analyzing our data, we observe a striking difference between Wikipedia articles and other genres regarding the distribution of SE-types (cf. Figure 7). For the selected English Wikipedia texts, 50% of the SE-types are GENERIC SENTENCE clauses, with STATES second at 24.3%.[8] 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 distribution of SE-types also differs according to genre: in argumentative texts, for example, GENERIC SENTENCES make up 48% of the SE-types, followed by STATES with a proportion of 32%, while in most other genres the most frequent class is STATE.

8. The Wikipedia texts were selected by Friedrich *et al.* (2015) precisely in order to target GENERIC SENTENCE clauses.

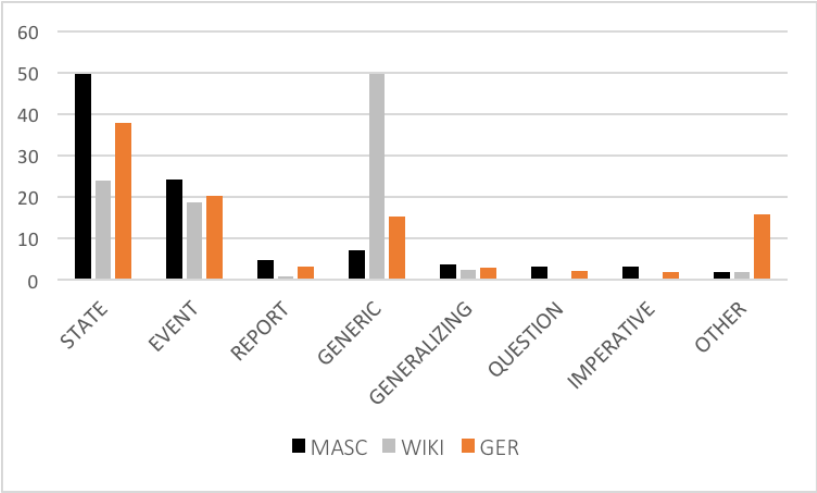

Figure 7. *Distribution of SE-types within different textual genres included in the corpus (in percentage). MASC=English corpus consisting of 12 genres; WIKI=English corpus consisting of Wikipedia articles, GER=German corpus consisting of 7 different genres.*

We also find that *sequences of SE-types* differ among genres: e.g., while STATE-STATE is the most frequent bigram within journal articles, the most frequent bigram in Wikipedia articles is GEN-GEN.[9] The most frequent trigram in jokes is EVENT-EVENT-EVENT, followed by STATE-STATE-STATE, whereas in government documents the most frequent trigrams are STATE-STATE-STATE and EVENT-STATE-STATE. In Section 7, we will further analyze such differences in the distribution of SE-label sequences across genres and how these differences are reflected by the performance of different model types.

5.3. *Confusability of Classes*

Some SE-types capture subtle aspectual distinctions, and as such are easily confused with one another. Friedrich (2017) (p. 93) provides a detailed analysis of annotator coincidence for the two English datasets used here. The three most inconsistently labeled types are STATES, GENERIC SENTENCES, and GENERALIZING SENTENCES. GENERALIZING SENTENCES are often labeled as EVENT or GENERIC SENTENCE; STATES are often labeled as GENERIC SENTENCE; and the label STATE is often applied to clauses labeled by other annotators as GENERIC SENTENCE, GENERALIZING SENTENCE, or EVENT.

9. GEN abbreviates GENERIC SENTENCE.

	Dataset	Eval	Acc	F1
Palmer07	Brown data	test	53.1	-
Fried16, set A (CRF)	MASC+Wiki	test	69.8	63.9
Fried16, set B (CRF)	MASC+Wiki	test	71.4	65.5
Fried16, set A+B (CRF)	MASC+Wiki	test	74.7	69.3
Fried16, set A+B (CRF)	MASC+Wiki	CV	76.4	71.2
Fried16, set A+B (CRF, seq-oracle)	MASC+Wiki	CV	77.9	73.9
BoW + SVM	MASC+Wiki	test	64.8	47.3

Table 2. *Reported results of baseline models for English from Palmer et al. 2007, Friedrich et al. 2016 and our own baseline system using word unigrams and bigrams as input for a SVM classifier (accuracy and macro-average F1 score). CV=10-fold cross-validation, test = evaluation on test set (20% of dataset, distinct set of documents in train and test). Since CV splits are not available, we only compare to the results on the held-out test set.*

5.4. *Segmentation*

The texts of the English dataset have been split into clauses using SPADE (Soricut and Marcu, 2003) with some heuristic post-processing. For the German dataset, DiscourseSegmenter's rule-based segmenter (EDSEG, Sidarenka *et al.* (2015)) was used. It uses German-specific rules to determine the boundaries of elementary discourse units in texts. Because DiscourseSegmenter occasionally oversplits segments, a small amount of post-processing was performed.

6. Experiments and Evaluation

6.1. *Baseline Systems*

From earlier work, the feature-based system of Palmer *et al.* (2007) (*Palmer07* in Table 2) simulates context through predicted labels from previous clauses. Their results are reported on 20 texts from the *popular lore* section of the Brown corpus (Francis and Kucera, 1979). Friedrich *et al.* (2016) (*Fried16* in Table 2) report results for their CRF-based SE-type labeler for different feature sets, evaluating both with 10-fold cross-validation and on a held-out test set (20% of the dataset, with distinct sets of documents in both train and test data). Training and testing were done on the combined MASC+Wiki dataset. *Fried16* is a sequence model which aims to learn the optimal global sequence of labels, jointly predicting labels for all clauses in a document. In the *oracle* setting, it includes the gold label of the previous clause. In our experiments, we adopt the models of *Fried16* as a very strong baseline for benchmarking our models, given that we are working on the same data.

Fried16's 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, as well as syntactic features, may not be captured in the word embeddings we use in our approach.

We also implement a simple baseline system which uses the CountVectorizer class from Sklearn (Pedregosa *et al.*, 2011) to calculate a bag of words matrix for the whole training and test data. Word unigrams and bigrams are used, and the resulting matrices are then fed into a LinearSVC classifier with default parameters.

Table 2 shows that, while *Palmer07* achieve modest results on Brown data, our BoW+SVM baseline is clearly lower than either of the feature-based CRF models of *Fried16* (on test set or 10-fold cross-validation setup). *Fried16's* results show that, when using sets A and B individually, Set B performs better than Set A on held-out test set, while their combination increases performance up to 74.7 accuracy. The result for the cross-validation setup (using feature set A+B) is very close to the seq-oracle result which includes the gold label of the previous clause. *Fried16* don't report seq-oracle results for the held-out test set. Please note that a direct comparison of our results to the results of the cross-validation setup of *Fried16* is not possible.

6.2. *Model Implementation and Training Setup*

Model implementation. We implemented the model variants for our SE-type classifier as described in Section 4, using Theano (Theano Development Team, 2016). We train the model with categorical cross entropy as loss function. For feature encoding (both SE labels in the context window and the genre label), we used 10-dimensional embedding vectors that we initialized randomly.

Test-train split. For the English dataset, we use the same test-train split as Friedrich *et al.* (2016). [10] 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, with 20% of training used for development (cf. Table 1). We report results in terms of accuracy and macro-average F1 score on the held-out test set.

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 LOC, 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).

10. The cross-validation splits of the data used by Friedrich *et al.* (2016) are not available.

For LOC_ATT the parameters are identical except for L2 (0.0001). We then apply the same hyperparameters as for LOC_ATT to the local model LOC_ATT+GEN and to the context models CON_TOK+GEN, CON_LAB+GEN and CON_LABTOK+GEN.

Window size as hyper-parameter. In the setting which includes previous labels we observe that the larger the window, the higher the accuracy. The opposite is the case for the context model which includes the *tokens* of previous clauses. We achieve best results when incorporating *five* previous clause labels in the CON_LAB* models or the tokens of *a single* previous clause in the CON_TOK* model (cf. Table 3). This also holds when porting our system to German (cf. Table 5). Adding more than five previous labels (or clauses) doesn't improve the system further.

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 syntactic and semantic 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 initialization which we trace back to the fact that our training data isn't large enough to derive good word embeddings. For German, we use 100-dimensional word2vec embeddings trained on a large German corpus of 116 million sentences using Skip-Gram mode with 5 negative samples (Reimers *et al.*, 2014). [11] For English, we use 300-dimensional word2vec embeddings (Mikolov *et al.*, 2013a) trained on a portion of the Google News dataset (about 100 billion words). The pre-trained embeddings are tuned during training. [12]

Testing for significance. To test significance of differences in accuracy, we apply McNemar's test with $p < 0.05$ and $p < 0.01$ rejecting the null hypothesis. Here we report significant differences between the best models (based on accuracy) for each of the four categories: local models, context models using tokens of previous clauses, context models using labels of previous clauses, and context models using both tokens and labels of previous clauses. When reporting the results, a pair of models that are significantly different from each other will be marked with the same symbol respectively for $p < 0.05$ and $p < 0.01$.

6.3. *Results*

Evaluation. We present the performance of the different models proposed in Section 4 in Table 3, reporting accuracy and macro F1 score on the test set. LOC achieves an accuracy of 66.55. Adding *attention* (LOC_ATT) yields an improvement of 2.63

11. https://public.ukp.informatik.tu-darmstadt.de /reimers/2014_german_embeddings.

12. We also experimented with FastText embeddings (Joulin *et al.*, 2017). Those embeddings take into account the internal structure of words which is especially useful for morphologically rich languages. We ran the local models (LOC, LOC_ATT and LOC_ATT+GEN) with FastText embeddings and found that word2vec embeddings work slightly better.

percentage points (pp). Using both *attention and genre* information (LOC_ATT+GEN) leads to a 1.94 pp increase over the model that uses only attention (LOC_ATT). Adding **context information** beyond the local clause in the form of the embedding representations of the **tokens of previous clauses** (CON_TOK+GEN) improves the model slightly, and a smaller window size yields better results than a larger one. The best results of this model type are obtained with the model which uses only the tokens of one previous clause jointly with genre information (CON_TOK1+GEN), and where the attention mechanism is applied only to the GRU of the target clause to be classified (71.67% accuracy). Using context in the form of **predicted labels of previous clauses** (CON_LAB+GEN) also improves the model. The model which uses the predicted labels of five previous clauses together with genre information (CON_LAB5+GEN) – with the attention mechanism being applied to the GRU of the target clause to be classified and to the representation of the previous labels – is our best performing model in general and yields an accuracy of 72.04.

The results in Table 3 show that using context information in the form of **predicted labels of previous clauses and embeddings for the tokens of previous clauses** in the CON_TOKLAB+GEN model is not favorable: accuracy drops compared to CON_TOK+GEN and CON_LAB+GEN, which use these two sources of contextual information separately.

Comparison to the CRF baseline model. All of our models outperform our simple baseline system BOW BL which uses word unigrams and bigrams as input for a SVM classifier. Both the CON_TOK1+GEN model and the CON_LAB5+GEN model outperform Friedrich *et al.* (2016)'s results both for the model that uses standard NLP features (feature set A) and the model that uses the more refined feature set B in isolation (cf. Table 2). Our models also come close to Friedrich *et al.'s* best results, which they obtain by applying their entire set of features including information from resources like WordNet, with a difference of 2.7 pp accuracy for our best performing model CON_LAB5+GEN. [13]

Attention vectors as input to sequence labeling models. 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 are then used in a conditional random field system (CRF++ [14].). The best performance was obtained when using the attention vector of the target 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. Future work includes the development of a CRF that can use continuous values.

Results for single classes. Figure 8 shows macro-average F1 scores of our best performing system CON_LAB5+GEN for the single SE classes. The scores are very similar to the results of Friedrich *et al.* (2016).

13. Since we did not reimplement their system, we cannot report significance results.
14. https://taku910.github.io/crfpp/

Model type		Model Name	Description	Acc	F1
Baselines	**BOW**	BOW+SVM	Bag of Words + SVM	64.83	47.3
	F+16 CRF	CRF, Set A	Standard feature set A	69.8	63.9
		CRF, Set B	Special SE feature set B	71.4	65.5
		CRF, Set A & B	Feature set A & B	**74.7**	**69.3**
Local		LOC	w/o attention	66.55	59.14
		LOC_ATT	with attention	69.18	68.31
		LOC_ATT+GEN	with attention + genre	**71.12**$^{\diamond\triangleleft\square\odot}$	**69.55**
Context	**Tokens**	CON_TOK1+GEN	1 prev. clause + genre	**71.67**$^{\diamond\odot}$	**59.19**
		CON_TOK2+GEN	2 prev. clauses + genre	71.57	48.12
		CON_TOK3+GEN	3 prev. clauses + genre	69.76	42.73
		CON_TOK4+GEN	4 prev. clauses + genre	69.29	41.55
		CON_TOK5+GEN	5 prev. clauses + genre	68.99	30.78
	Labels	CON_LAB1+GEN	1 prev. label + genre	69.55	60.21
		CON_LAB2+GEN	2 prev. labels + genre	71.04	64.54
		CON_LAB3+GEN	3 prev. labels + genre	71.68	64.42
		CON_LAB4+GEN	4 prev. labels + genre	71.25	**65.06**
		CON_LAB5+GEN	5 prev. labels + genre	**72.04**$^{\triangleleft}$	64.74
	Tokens + Labels	CON_TOKLAB1+GEN	1 prev. label/clause + genre	**71.35**$^{\square}$	**70.82**
		CON_TOKLAB2+GEN	2 prev. labels/clauses + genre	70.65	68.62
		CON_TOKLAB3+GEN	3 prev. labels/clauses + genre	69.90	68.83
		CON_TOKLAB4+GEN	4 prev. labels/clauses + genre	69.26	67.47
		CON_TOKLAB5+GEN	5 prev. labels/clauses + genre	69.00	64.36

Table 3. *SE-type classification on English test set. For our models using context information, we only report the results for the best performing models for the following three categories: (i) context models using tokens of previous clauses (*CON_TOK+GEN*); (ii) context models using labels of previous clauses (*CON_LAB+GEN*); and (iii) context models using tokens and labels of previous clauses jointly (*CON_TOKLAB+GEN*). F1 is reported as macro-average score. Significance based on accuracies is computed for the best performing models of each category; pairs of models that are significantly different from each other share the same symbol. Models with the symbol $\odot$ are also signficant with $p < 0.01$.*

Scores for GENERALIZING SENTENCE are the lowest as this class is very infrequent in the data set, while scores for the classes STATE, EVENT, QUESTION and OTHER are the highest. In addition, we explored system performance of CON_LAB5+GEN in a binary (one vs. rest, OvR) classification setting, classifying STATE vs. the remaining classes, EVENT vs. the remaining classes, etc. (cf. Figure 8). Binary classification achieves better performance and can be useful for other tasks which only need information about specific SE-types, for example for distinguishing generic from non-generic sentences. Becker *et al.* (2017) for example showed,

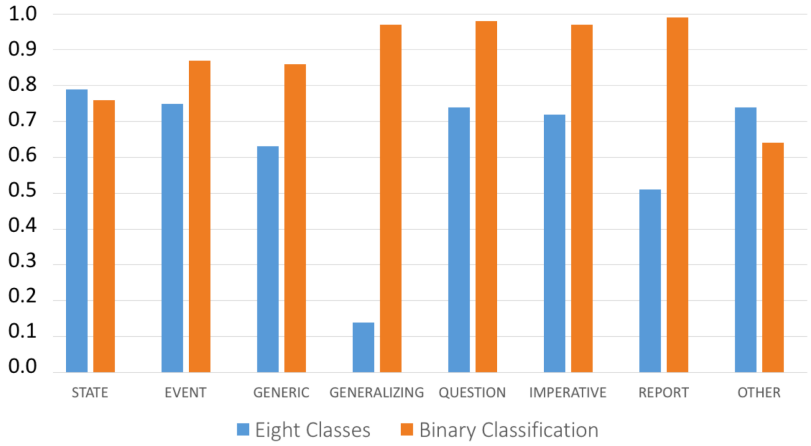

Figure 8. *Macro-average F1 scores of our best performing system* CON_LAB5+GEN *for single SE classes, multiclass vs. binary classification.*

in an annotation experiment of linguistic characteristics of implicit knowledge within argumentative texts, that a majority of implicit information is encoded as GENERIC SENTENCES. This tendency could be deployed for acquiring such knowledge automatically. Other possible application for binary classifications would be extracting non-canonical imperatives or questions for dialogue systems or distinguishing events from non-events as part of systems for event extraction or veridicality determination.

Confusability of classes. Section 5.3 discusses the SE-types with high confusability for human annotators. Most of these same confusions occur with high frequency in the outputs from our best model, as shown in Table 4. In particular, GENERALIZING SENTENCES are often mislabeled as EVENT or STATE – e.g., the clause *Even friendly nations routinely steal information from US companies* is a GENERALIZING SENTENCE but gets labeled as STATE by our model. A major reason for frequently mislabeling GENERALIZING SENTENCES could be that this class is very small within in our dataset (cf. Fig 7). We also find that STATES and GENERIC SENTENCES are frequently confused: e.g., the clause *Certainly the Colombian press is much in need of that* is labeled as GENERIC SENTENCE by our classifier, while the gold label is STATE.

7. Impact and Analysis of Attention and Genre

Our experimental results in Section 6 show that both attention and incorporation of genre information result in improved performance for our local models. In this section, we look more closely at the role of these factors in SE classification.

	Event	Report	State	GenSt	Generic	IMP	Q	Other
Event	**1,255**	25	249	8	58	7	5	115
Report	43	**243**	14	-	-	-	1	16
State	224	6	**2,912**	23	219	31	15	125
GenSt	87	3	109	**40**	58	8	2	19
Generic	70	1	446	12	**948**	8	2	91
IMP	7	-	19	1	7	**165**	1	35
Q	8	1	69	-	3	8	**96**	23
Other	171	2	216	16	117	29	8	**1,306**

Table 4. *Confusion matrix for English of our best performing model* CON_LAB5+GEN.

7.1. *Analysis of Attention*

Attention is an effective mechanism that allows models to focus on specific parts of the input instead of capturing the complete semantics of the input in its hidden state. Beyond this capacity, attention can also give insights into what elements the model learns to be most relevant for predicting the various SE-types. The analyses reported in this section are based on the output of LOC_ATT+GEN, our best performing local model, which is run on the English dataset and uses as input the target clause to be classified jointly with attention and genre information.

We analyze the attention weights learned by the model and focus on different linguistic information: (1) The attention to specific words for specific SE-types; (2) the attention to specific POS tags for specific SE-types and 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.

Attention to specific words. When analyzing the characteristics of SE-types regarding words which are assigned high attention scores during training, we find that different classes of words are highlighted for different SE-types. For STATES, nouns and personal pronouns (*youngsters, editors, joyce, I, me*) as well as predicative auxiliaries (*am, are, is*) play a predominant role. In clauses classified as EVENTS, we find many gerunds (*thinking, writing*) with high attention scores, while for GENERIC SENTENCES, adjectives and adverbs (*chronic, awake*), modal verbs (*can, may, must*) and indefinite determiners (*a, an*) are given high attention scores. Interestingly, we find many named entities with high attention scores (*york, states, Miller*) when classifying GENERALIZING SENTENCES. High attention scores for predicative auxiliaries when classifying STATES or for gerunds when classifying EVENTS make sense linguistically. Modal verbs as indicators for GENERIC SENTENCES are very motivated as well, as we often find them with assertions over kinds, such as *Birds can fly, Children must go to school.* However, other findings (e.g., the predominant role of nouns for STATES or of adjectives and adverbs for GENERIC SENTENCES) seem to be arbitrary.

Next, we focus on particular clauses for which adding attention leads to improved classification. Here we analyze attention scores only for those instances that are classified correctly by the attention-enhanced model, while they are incorrectly classified by the model without attention. In these instances, we find many verbs, in particular verbs in past tense (*helped, submitted, included*), which are assigned high attention scores within clauses classified as STATES. Within EVENTS, discourse markers and modifiers (*well, but, some*) are given high attention scores, while in GENERIC SENTENCES modal verbs (*allows, can, must*) play a predominant role. Finally, verba dicendi such as *reported, explained, tells,* or *cited*, get high attention scores when classifying REPORTS, while for the correct classification of QUESTIONS, interrogatives (*what, when, where*) are important. These observations mostly make sense linguistically and highlight the key role of certain word classes.

Attention to specific POS tags. We complement the analysis of attention to specific words with a systematic analysis of attention to POS tags. We therefore post-process our data with POS tags using *spaCy* [15] with the Penn Treebank Tagset (Marcus *et al.*, 1993). Figure 9 visualizes the mean attention score per POS tag for all SE-types (gold labels).

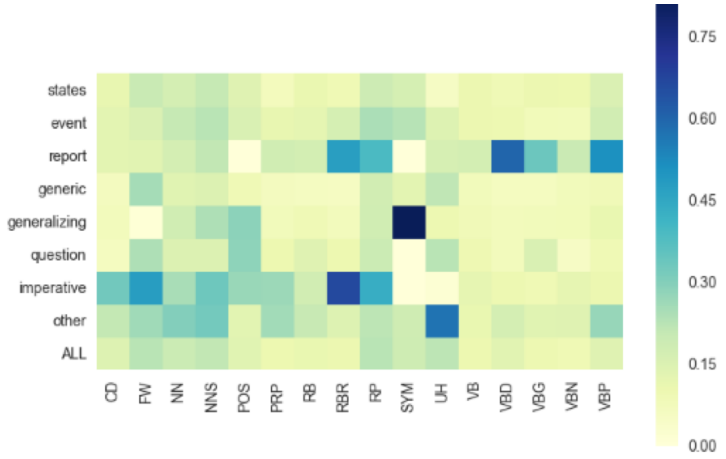

Figure 9. *Mean attention scores for specific SE-types per POS tag on the English training set. POS tags from PTB.*

Interestingly, when we move from focusing on words to focusing on the much broader categories of POS classes, attention weights stand out for classes that are rare, such as IMPERATIVE or REPORT, each less than 5% of the English dataset. Thus, in contrast to sparse lexical material, attention here seems to focus on some more abstract word properties. We don't find outstanding attention weights for particular POS tags when classifying frequent SE-types such as STATE, EVENT or GENERIC SENTENCE.

15. https://spacy.io/docs/usage/pos-tagging.

The heat map indicates that the model attends especially to verbs when classifying the SE-type REPORT. This is not surprising, since REPORT clauses such as *he said* are signaled by verbs of speech. GENERALIZING SENTENCES 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*.

Position of words with high attention scores. Figure 10 shows the relative positions of words with maximum and high 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. Thus, the model seems to draw information from a broad range of positions.

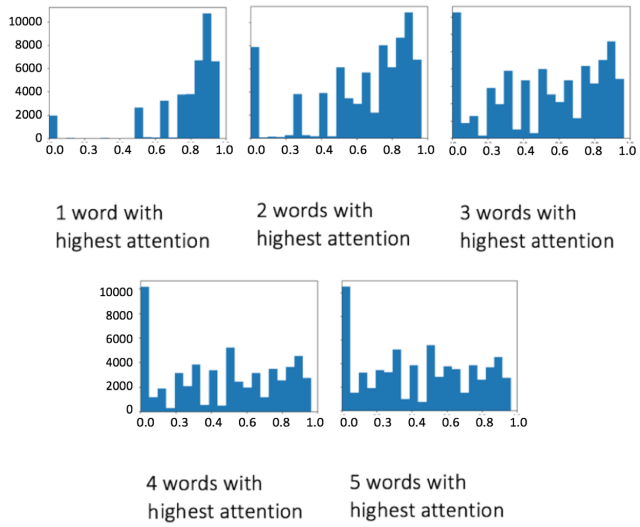

Figure 10. *Position of words with maximum attention within clauses; x-axis represents the normalized position within the clause, y-axis the number of words with maximum attention at that position.*

The analysis of attention weights yields a number of expected findings, mostly for easily-characterized (though infrequent) classes such as REPORT and QUESTION. For more frequent and more varied classes such as EVENT, STATE, and GENERIC SENTENCE, neither single words nor single POS tags seem to provide especially strong signals for SE classification. Analysis of attention and word position indeed suggests that the model attends to multiple elements of the clause in order to arrive at an SE-label.

7.2. *Analysis of Genre*

We investigate more deeply the relevance of genre information on the classification performance and to what extent genre information is reflected in SE-label sequences. [16]

We first compare the accuracy of the best performing local model and the best performing context model, respectively with and without genre information (LOC_ATT+GEN vs. LOC_ATT and CON_LAB5+GEN vs. CON_LAB5), for the English dataset. The results are given in Figure 11. For some genres (e.g., news, ficlets, and emails), LOC_ATT performs quite well and shows little benefit from the inclusion of genre information. For governmental protocols, technical reports, Wikipedia articles, travel reports, and letters, on the other hand, LOC_ATT benefits quite a bit from genre information (i.e., LOC_ATT+GEN far outperforms LOC_ATT). CON_LAB5, which uses context in form of the labels of previous clauses, in general seems not to benefit as much from genre information as LOC_ATT (while performing better overall due to context information). Figure 11 shows for example that genres such as Ficlets, Emails, Fiction or Journal articles benefit very little or not at all from the inclusion of genre information, while governmental protocols and letters benefit from genre information.

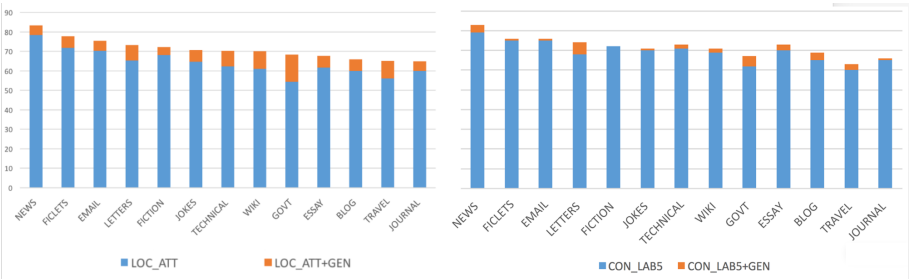

Figure 11. *Comparing performances (accuracy) of our best performing local model* LOC_ATT *and our best performing context model* CON_LAB5*, respectively with and without genre information, for the genres of our English test set separately.*

In order to further analyze the effects of genre information on SE-type classification, we explore the similarity of genres with respect to the distribution of SE-types

16. We didn't train the classifier strictly within specific genres mainly for two reasons: The first reason is that we have too little data for some genres such as technical reports, letters or essays (see Figures 11 and 12), and even for the genres with a comparably high number of instances such as Wikipedia, the data size is still quite low to train our system sufficiently without overfitting. But even more important, one crucial aim in developing the classifier was to build a system which is robust across genres when classifying SE-types, which highlights the importance of training our model across various genres at the same time.

and their sequences (modeled as bigrams) measured by symmetric Kullback-Leibler divergence:

$$D_{klsym}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \quad [12]$$

where

$$D_{KL}(P||Q) = \sum_i P(i) * \log \frac{P(i)}{Q(i)} \quad [13]$$

P and Q are the corresponding distributions of SE-types (unigrams or bigrams) for different genres, whereas i iterates over all possible outcomes. The results based on unigrams and bigrams of SE-types are visualized as a heat map in Figure 12. News and Wikipedia articles as well as news and emails show high values and therefore differ a lot with respect to the distribution of both uni- and bigrams of SE-types, while jokes and blog articles or journal articles and fiction are more similar. For some genres (Wikipedia articles in particular) the findings from this analysis correspond to the size of performance improvement due to incorporation of genre information (cf. Figure 11). For others (e.g., news and emails) the correspondence doesn't hold.

Figure 12. *Distributional divergence of SE-types across genres: symmetric Kullback-Leibler divergence of unigrams (left) and bigrams (right) of SE-types within the English dataset (train+dev+test set).*

Overall, bigrams show larger differences than unigrams. Wikipedia articles, followed by travel reports, show the highest values both for unigram and bigram analyses. We also see that Wikipedia articles and travel reports are both dissimilar to essays, ficlets, fiction, and news, while essays, ficlets, and fiction are on the other hand quite similar to government documents (light green). This suggests that the improvements in Wikipedia articles and travel reports are strongly related to each other, and that these genres profit mutually from the genre information given to the model. The distributional similarities and dissimilarities seem to be exploited in the models using genre information. We can also expect that genres that are sparse can profit from genres with

Model type		Name	Description	Acc	F1
Local		LOC	w/o attention	74.94	67.12
		LOC_ATT	with attention	74.51	**74.02**
		LOC_ATT+GEN	with attention + genre	**75.56**$^{\diamond\triangleleft\square}$	69.98
Context	**Tokens**	CON_TOK1+GEN	1 prev. clause + genre	**74.51**$^{\diamond}$	**72.41**
		CON_TOK2+GEN	2 prev. clauses + genre	74.44	72.26
		CON_TOK3+GEN	3 prev. clauses + genre	73.35	71.79
		CON_TOK4+GEN	4 prev. clauses + genre	73.11	71.12
		CON_TOK5+GEN	5 prev. clauses + genre	72.89	70.61
	Labels	CON_LAB1+GEN	1 prev. label + genre	71.78	**52.88**
		CON_LAB2+GEN	2 prev. labels + genre	72.29	52.52
		CON_LAB3+GEN	3 prev. labels + genre	72.47	52.34
		CON_LAB4+GEN	4 prev. labels + genre	74.33	51.12
		CON_LAB5+GEN	5 prev. labels + genre	**74.92**$^{\triangleleft}$	50.76
	Tokens + Labels	CON_TOKLAB1+GEN	1 prev. label/clause + genre	**73.43**$^{\square}$	**59.51**
		CON_TOKLAB2+GEN	2 prev. labels/clauses + genre	72.23	57.38
		CON_TOKLAB3+GEN	3 prev. labels/clauses + genre	71.69	57.99
		CON_TOKLAB4+GEN	4 prev. labels/clauses + genre	71.11	56.48
		CON_TOKLAB5+GEN	5 prev. labels/clauses + genre	71.09	56.23

Table 5. *SE-type classification on German test set. Again, we only report the results for the best performing models for the context models (*CON_TOK+GEN*,* CON_LAB+GEN *and* CON_TOKLAB+GEN*). F1 is reported as macro-average score. Pairs of models that yield significant performance differences are marked with the same symbol; significance is computed for the best performing models of each category.*

similar SE-type distributions that are more frequent (cf. Figure 6). In future work, it would be interesting to consider other approaches to measuring genre similarity, such as overlap of lexical items, syntactic structures, or topic model distributions.

8. Porting the System to German

A great advantage of neural-based systems is that they are able to learn relevant features for classification during the training procedure, and thus do not rely on hand-crafted features. This is of considerable help when models are to be transferred to novel languages, when such features are expensive to compute, or not available because of lack of resources.

We use the system described above with German data, and adjust the size of the input embeddings. [17] We tune hyperparameters separately for German on the German development set through random search using *optunity* (Bergstra and Bengio, 2012) and use the best setting for evaluating on the test set. As for the English dataset, we tune batch size, number of layers, GRU cell size, and regularization parameter (L2), we use AdaGrad (Duchi *et al.*, 2011) for learning rate optimization, and we tune the initial learning rate. For LOC, the best result on the development set is achieved for GRU with batch size 100, 2 layers, cell size 124, learning rate 0.05, and L2 regularization parameter (0.01). For LOC_ATT the parameters are identical, except for L2 (0.0001). As for English, we again apply the optimal hyperparameters for LOC_ATT to the local model LOC_ATT+GEN and to the context models CON_TOK+GEN, CON_LAB+GEN and CON_LABTOK+GEN.

Table 5 gives an overview of the results for different models, and allows us to compare the effectiveness of integrating context and genre information. Compared to English, the local models achieve higher performance, but attention by itself does not improve the results (cf. LOC vs. LOC_ATT). Used jointly, attention and genre information LOC_ATT+GEN yield a moderate increase of 0.62 pp accuracy compared to LOC. Attention may need more data and possibly more diversity to be learned effectively. Improving the attention models for German will be the focus of our future work, to facilitate a meaningful linguistic analysis of attention for German.

In the case of German, modeling context information doesn't improve results: compared to the best local model (LOC_ATT+GEN), adding more context either in form of the tokens (CON_TOK+GEN) or labels of previous clauses (CON_LAB+GEN) or both (CON_TOKLAB+GEN) does not lead to higher accuracy (cf. Table 5). Again, this can be due to the smaller dataset size, which may not provide enough data points for the richer context models.

9. Conclusion

We presented an RNN-based approach to SE-type classification that bears clear advantages compared to previous classifier models that rely on sophisticated, hand-engineered features and lexical semantic resources: given pre-annotated training data, our neural model is easily transferable to other languages as it can tune pre-trained word embeddings to encode semantic information relevant for the task.

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 exploit 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

17. The different size of the embeddings (for English and German cf. Section 4.4) may have an impact on the results.

from the contextual history, and the analysis of the interaction between these properties and genre characteristics as well as the interaction of sequence information and genre.

Our best model trained on English data jointly uses genre and context information in the form of previously predicted labels and is enhanced with attention. It outperforms the state-of-the-art models of Friedrich *et al.* (2016) 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 less than 3 pp accuracy is left to achieve in future work to compete with the knowledge-rich model combining both feature sets.

For the German models we find that the local model enhanced with attention and genre information leads to highest accuracies, while modeling context information doesn't improve the results. We leave improving the attention and context models for German as future work.

For our English dataset, we show that, by using attention, we can gain insights into what the models learn. The analysis of attention weights shows interesting findings, especially for easily characterized classes such as REPORT and QUESTION, while for more varied classes such as EVENT, STATE, and GENERIC SENTENCE, we don't observe clear patterns or distributions regarding single words or POS tags (which could be helpful for classification), notwithstanding the fact that the attention mechanism in general improves our models. Some of our findings can be motivated linguistically, e.g., high attention scores for predicative auxiliaries when classifying STATES or modal verbs as indicators for GENERIC SENTENCES. We further observe that our models attend to multiple elements of the clause during training and therefore seem to draw information from a broad range of positions within clauses.

Our analyses of the impact of genre information (again on the English dataset) show that genre improves classification performance across the board, but some genres benefit more from this information than others, which can be partially linked to variation in SE-type distributions. Our models can be used either as multi-class or as binary classifiers for detecting events, generics, imperatives, or questions; they can help model discourse modes or can improve argument analysis and argument detection tasks.

Acknowledgments. We thank Sabrina Effenberger, Jesper Klein, Sarina Meyer, and Rebekka Sons for the annotations and their helpful feedback on the annotation manual. 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. We also acknowledge a grant from NVIDIA Corporation.

10. References

Abdul-Mageed M., Ungar L., "EmoNet: Fine-Grained Emotion Detection With Gated Recurrent Neural Networks", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 718-728, 2017.

Asher N., *Reference to Abstract Objects in Discourse*, Kluwer Academic Publishers, 1993.

Bahdanau D., Cho K., Bengio Y., "Neural Machine Translation by Jointly Learning to Align and Translate", *International Conference on Machine Learning*, 2015.

Becker M., Palmer A., Frank A., "Argumentative Texts and Clause Types", *Proceedings of the 3rd Workshop on Argument Mining*, p. 21-30, 2016a.

Becker M., Palmer A., Frank A., "Clause Types and Modality in Argumentative Microtexts", *Workshop on Foundations of the Language of Argumentation (in conjunction with COMMA)*, p. 1-9, 2016b.

Becker M., Staniek M., Palmer A., Nastase V., Frank A., "Classifying Semantic Clause Types: Modeling Context and Genre Characteristics with Recurrent Neural Networks and Attention", *Proceedings of the Joint Conference on Lexical and Computational Semantics (Starsem)*, p. 230-240, 2017.

Bengio Y., Simard P., Frasconi P., "Learning Long-term Dependencies with Gradient Descent is Difficult", *IEEE Transactions of Neural Networks*, vol. 5, n° 2, p. 157-166, 1994.

Bergstra J., Bengio Y., "Random Search for Hyper-Parameter Optimization", *Journal of Machine Learning Research*, vol. 13, p. 281-305, 2012.

Brown P. F., Desouza P. V., Mercer R. L., Pietra V. J. D., Lai J. C., "Class-Based N-gram Models of Natural Language", *Computational Linguistics*, vol. 18, n° 4, p. 467–479, 1992.

Carlson G. N., Pelletier F. J. (eds), *The Generic Book*, University of Chicago Press, 1995.

Cheng F., Miyao Y., "Classifying Temporal Relations by Bidirectional LSTM over Dependency Paths", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 1-6, 2017.

Cho K., van Merrienboer B., Bahdanau D., Bengio Y., *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*, 8th Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014.

Conneau A., Kiela D., Schwenk H., Barrault L., Bordes A., "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 670-680, 2017.

Costa F., Branco A., "Aspectual Type and Temporal Relation Classification", *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 266-275, 2012.

Dai Z., Huang R., "Building Context-aware Clause Representations for Situation Entity Type Classification", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 3305-3315, 2018.

Dong L., Lapata M., "Language to Logical Form With Neural Attention", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 33-43, 2016.

Dowty D., *Word Meaning and Montague Grammar*, Reidel, 1979.

Duchi J., Hazan E., Singer Y., "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization", *Journal of Machine Learning Research*, vol. 12, n° 7, p. 2121-2159, 2011.

Francis N., Kucera H., "A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers.", *Department of Linguistics, Brown University, Providence, Rhode Island, USA*, 1979.

Friedrich A., States, Events, and Generics: Computational Modeling of Situation Entity Types, PhD thesis, Universität des Saarlandes, 2017.

Friedrich A., Palmer A., "Automatic Prediction of Aspectual Class of Verbs in Context", *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 517-523, 2014a.

Friedrich A., Palmer A., "Situation Entity Annotation", *Proceedings of the Linguistic Annotation Workshop VIII*, p. 149-158, 2014b.

Friedrich A., Palmer A., Pinkal M., "Situation Entity Types: Automatic Classification of Clause-Level Aspect", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 1757-1768, 2016.

Friedrich A., Palmer A., Sørensen M. P., Pinkal M., "Annotating Genericity: a Survey, a Scheme, and a Corpus", *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, p. 21, 2015.

Friedrich A., Pinkal M., "Discourse-sensitive Automatic Identification of Generic Expressions", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 1272-1281, 2015.

Goldberg Y., *Neural Network Methods for Natural Language Processing*, vol. 37 of *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool, San Rafael, CA, 2017.

Goodfellow I., Bengio Y., Courville A., *Deep Learning*, MIT Press, 2016. `http://www.deeplearningbook.org`.

Herbelot A., Copestake A., "Annotating Genericity: How Do Humans Decide? (A Case Study in Ontology Extraction)", *Studies in Generative Grammar*p. 103, 2009.

Hochreiter S., Schmidhuber J., "Long Short-Term Memory", *Neural computation*, vol. 9, n° 8, p. 1735-1780, 1997.

Ide N., Fellbaum C., Baker C., Passonneau R., "The Manually Annotated Sub-Corpus: A Community Resource For and By the People", *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 68-73, 2010.

Joulin A., Grave E., Bojanowski P., Mikolov T., "Bag of Tricks for Efficient Text Classification", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,*p. 427-431, 2017.

Kalchbrenner N., Grefenstette E., Blunsom P., "A Convolutional Neural Network for Modelling Sentences", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Baltimore, Maryland, p. 655-665, 2014.

Kim Y., "Convolutional Neural Networks for Sentence Classification", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, p. 1746-1751, 2014.

Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C., "Neural Architectures for Named Entity Recognition", *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, p. 260-270, 2016.

Loaiciga S., Meyer T., Popescu-Belis A., "English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling", *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, 2014.

Marcus M., Santorini B., Marcinkiewicz M. A., "Building a Large Annotated Corpus of English: The Penn Treebank", *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1993.

Mavridou K.-I., Friedrich A., Sorensen M., Palmer A., Pinkal M., "Linking Discourse Modes and Situation Entities in a Cross-Linguistic Corpus Study", *Proceedings of the EMNLP Workshop LSDSem 2015: Linking Models of Lexical, Sentential and Discourse-level Semantics*, 2015.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., "Distributed Representations of Words and Phrases and Their Compositionality", *Advances in neural information processing systems*, p. 3111-3119, 2013a.

Mikolov T., Yih W.-t., Zweig G., "Linguistic Regularities in Continuous Space Word Representations", *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, p. 746-751, 2013b.

Mishra A., Dey K., Bhattacharyya P., "Learning Cognitive Features from Gaze Data for Sentiment and Sarcasm Classification using Convolutional Neural Network", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 377-387, 2017.

Miwa M., Bansal M., "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 1105-1116, 2016.

Nedoluzhko A., "Generic Noun Phrases and Annotation of Coreference and Bridging Relations in the Prague Dependency Treebank", *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 103-111, 2013.

Nie A., Bennett E. D., Goodman N. D., "DisSent: Sentence Representation Learning from Explicit Discourse Relations", *Computing Research Repository*, 2017.

O'Seaghdha D., Teufel S., "Unsupervised Learning of Rhetorical Structure with Un-Topic Models", *Proceedings of the 25th International Conference on Computational Linguistics*, Association for Computational Linguistics, p. 2-13, 2014.

Palmer A., Friedrich A., "Genre Distinctions and Discourse Modes: Text Types Differ in Their Situation Type Distributions", *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and NLP*, 2014.

Palmer A., Ponvert E., Baldridge J., Smith C., "A Sequencing Model for Situation Entity Classification", *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 896-903, 2007.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M.,

Perrot M., Duchesnay E., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, 2011.

Peldszus A., Stede M., "An Annotated Corpus of Argumentative Microtexts", *Proceedings of the First European Conference on Argumentation*, 2015.

Plank B., Søgaard A., Goldberg Y., "Multilingual Part-of-Speech Tagging With Bidirectional Long Short-Term Memory Models and Auxiliary Loss", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 412-418, 2016.

Reimers N., Eckle-Kohler J., Schnober C., Kim J., Gurevych I., "Germeval-2014: Nested Named Entity Recognition with Neural Networks", *Proceedings of the 12th Edition of the KONVENS Conference*, p. 117–120, 2014.

Reiter N., Frank A., "Identifying Generic Noun Phrases", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, p. 40-49, 2010.

Rocktäschel T., Grefenstette E., Hermann K. M., Kočiskỳ T., Blunsom P., "Reasoning About Entailment With Neural Attention", *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, May, 2016.

Sanagavarapu K. C., Vempala A., Blanco E., "Determining Whether and When People Participate in the Events They Tweet About", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 641-646, 2017.

Sidarenka U., Peldszus A., Stede M., "Discourse Segmentation of German Texts", *Journal for Language Technology and Computational Linguistics*, vol. 30, p. 71-98, 2015.

Siegel E. V., McKeown K. R., "Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights", *Computational Linguistics*, vol. 26, n° 4, p. 595-628, 2000.

Smith C. S., *The Parameter of Aspect*, Kluwer, 1991.

Smith C. S., *Modes of Discourse: The Local Structure of Texts*, vol. 103, Cambridge University Press, 2003.

Smith C. S., "Aspectual Entities and Tense in Discourse", *Aspectual inquiries*, Springer, p. 223-237, 2005.

Song W., Wang D., Fu R., Liu L., Liu T., Hu G., "Discourse Mode Identification in Essays", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 112-122, 2017.

Soricut R., Marcu D., "Sentence Level Discourse Parsing Using Syntactic and Lexical Information", *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.*, Association for Computational Linguistics, p. 149-15, 2003.

Tai K. S., Socher R., Manning C. D., "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, p. 1556-1566, 2015.

Teufel S., Argumentative Zoning: Information Extraction From Scientific Text, PhD thesis, University of Edinburgh, 2000.

Theano Development Team, "Theano: A Python Framework for Fast Computation of Mathematical Expressions", *arXiv e-prints*, May, 2016.

Vempala A., Blanco E., Palmer A., "Determining Event Durations: Models and Error Analysis", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, p. 164-168, 2018.

Vendler Z., "Verbs and Times", *The Philosophical Review,*p. 143-160, 1957.

Verkuyl H., *On the Compositional Nature of the Aspects*, Reidel, 1972.

Vu N. T., Adel H., Gupta P., Schütze H., "Combining Recurrent and Convolutional Neural Networks for Relation Classification", *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 534-539, 2016.

Wang Y., Huang M., zhu x., Zhao L., "Attention-based LSTM for Aspect-level Sentiment Classification", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 606-615, 2016.

Yin W., Kann K., Yu M., Schütze H., "Comparative Study of CNN and RNN for Natural Language Processing", *Computing Research Repository*, 2017.

Zarcone A., Lenci A., "Computational Models of Event Type Classification in Context", *Proceedings of the International Conference on Language Resources and Evaluation*, 2008.

Zhou J., Cao Y., Wang X., Li P., Xu W., "Deep Recurrent Models With Fast-Forward Connections for Neural Machine Translation", *Transactions of the Association for Computational Linguistics,*p. 371-383, 2016.

Zhou J., Xu W., "End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, p. 1127-1137, 2015.