Situation entity types: automatic classification of clause-level aspect

Annemarie Friedrich[1] **Alexis Palmer**[2] **Manfred Pinkal**[1]

[1]Department of Computational Linguistics, Saarland University, Germany
[2]Leibniz ScienceCampus, Dept. of Computational Linguistics, Heidelberg University, Germany
{afried,pinkal}@coli.uni-saarland.de palmer@cl.uni-heidelberg.de

Abstract

This paper describes the first robust approach to automatically labeling clauses with their situation entity type (Smith, 2003), capturing aspectual phenomena at the clause level which are relevant for interpreting both semantics at the clause level and discourse structure. Previous work on this task used a small data set from a limited domain, and relied mainly on words as features, an approach which is impractical in larger settings. We provide a new corpus of texts from 13 genres (40,000 clauses) annotated with situation entity types. We show that our sequence labeling approach using distributional information in the form of Brown clusters, as well as syntactic-semantic features targeted to the task, is robust across genres, reaching accuracies of up to 76%.

1 Introduction

Clauses in text have different aspectual properties, and thus contribute to the discourse in different ways. Distinctions that have been made in the linguistic and semantic theory literature include the classification of states, events and processes (Vendler, 1957; Bach, 1986), and whether clauses introduce particular eventualities or report regularities generalizing either over events or members of a kind (Krifka et al., 1995). Such aspectual distinctions are relevant to natural language processing tasks requiring text understanding such as information extraction (Van Durme, 2010) or temporal processing (Costa and Branco, 2012).

In this paper, we are concerned with automatically identifying the **type** of a **situation entity (SE)**, which we assume to be expressed by a clause. Specifically, we present a system for automatically labeling clauses using the inventory of SE types shown in Figure 1 (Smith, 2003, 2005; Palmer et al., 2007). The original motivation for the above inventory of SE types is the observation that different *modes of discourse*, a classification of linguistic properties of text at the passage level, have different distributions of SE types. For example, EVENTs and STATEs are predominant in *narrative* passages, while GENERIC SENTENCEs occur frequently in *information* passages.

STATE: *The colonel owns the farm.*
EVENT: *John won the race.*
REPORT: "...", *said Obama.*
GENERIC SENTENCE: Generalizations over kinds. *The lion has a bushy tail.*
GENERALIZING SENTENCE: Generalizations over events (*habituals*). *Mary often fed the cat last year.*
QUESTION: *Who wants to come?*
IMPERATIVE: *Hand me the pen!*

Figure 1: SE types, adapted from Smith (2003).

A previous approach to automatically labeling SE types (Palmer et al., 2007) – referred to here as UT07 – captures interesting insights, but is trained and evaluated on a relatively small amount of text (about 4300 clauses), mainly from one rather specialized subsection of the Brown corpus. The data shows a highly skewed distribution of SE types and was annotated in an intuitive fashion with only moderate agreement. In addition, the UT07 system relies mostly on part-of-speech tags and words as features. The latter are impractical when dealing with larger data sets and capture most of the corpus vocabulary, overfitting the model to the data set. Despite this overfitting, the system's accuracy is only around 53%.

We address these shortcomings, developing a robust system that delivers high performance compared to the human upper bound across a range of genres. Our approach uses features which increase

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1757–1768,
Berlin, Germany, August 7-12, 2016.

robustness: Brown clusters and syntactic-semantic features. Our models for labeling texts with the aspectual properties of clauses in the form of SE types reach accuracies of up to 76%.

In an oracle experiment, Palmer et al. (2007) show that including the gold labels of the previous clauses as features into their maximum entropy model is beneficial. We implement the first true sequence labeling model for SE types, using conditional random fields to find the globally-best sequence of labels for the clauses in a document. Performance increases by around 2% absolute compared to predicting labels for clauses separately; much of this effect stems from the fact that GENERIC SENTENCEs often occur together.

Moving well beyond the single-domain setting, our models are trained and evaluated on a multi-genre corpus of approximately 40,000 clauses from MASC (Ide et al., 2008) and Wikipedia which have been annotated with substantial agreement. We train and test our models both within genres and across genres, highlighting differences between genres and creating models that are robust across genres. Both the corpus and the code for an SE type labeler are freely available.[1] These form the basis for future research on SE types and related aspectual phenomena and will enable the inclusion of SE type information as a preprocessing step into various NLP tasks.

2 Linguistic background

The inventory of SE types proposed by Smith (2003) consists of three high-level categories, each with two subtypes. *Eventualities* comprise EVENT and STATE, categories for clauses representing actual happenings, states of the world, or attributes of entities or situations. *General Statives* include GENERIC SENTENCE and GENERALIZING SENTENCE and reflect regularities in the world or general information predicated over classes or kinds. Finally, *Abstract Entities* (Figure 2) have the subtypes FACT and PROPOSITION. Although *Abstract Entities* are part of the label inventory for UT07, we treat them in a separate identification step, for reasons discussed in Section 7. The inventory was expanded by Palmer et al. (2007) to include three additional types: REPORT, QUESTION and IMPERATIVE. The latter two categories were added to accommodate exhaustive annotation of text; REPORT is a subtype of event for attributions of quoted speech.

FACT: Objects of knowledge.
I know that Mary refused the offer.
PROPOSITION: Objects of belief.
I believe that Mary refused the offer.

Figure 2: *Abstract Entity* SE types.

Two parts of a clause provide important information for determining the SE type (Friedrich and Palmer, 2014b): a clause's **main verb** and its **main referent**. The latter is loosely defined as the main entity that the segment is about; in English this is usually the subject. For example, main referents of GENERIC SENTENCEs are kinds or classes as in "*Elephants are huge*", while the main referents of *Eventualities* and GENERALIZING SENTENCEs are particular individuals ("*John is short*"). For English, the main verb is the non-auxiliary verb ranked highest in the dependency parse (e.g. "*kiss*" in "*John has kissed Joe*"). STATEs and EVENTs differ in the *fundamental lexical aspectual class* (Siegel and McKeown, 2000) of their main verbs (e.g. **dynamic** in "*She filled the glass with water*" vs. **stative** in "*Water fills the glass*"). While fundamental lexical aspectual class is a word-sense level attribute of the clause's main verb, *habituality* is a property of the entire clause which is helpful to determine the clause's SE type. For example, EVENT and GENERALIZING SENTENCE differ in habituality (e.g. **episodic** in "*John cycled to work yesterday*" vs. **habitual** in "*John cycles to work*").

Like habituality, SE types are a categorization at the clause level. Properties of the clause such as modals, negation, or the perfect influence the SE type: for instance, "*John might win*" is treated as a STATE as it describes a possible state of the world rather than an EVENT. Such coercions happen only for clauses which, without the trigger for aspectual shift, would be EVENTs; other SEs retain their type even under coercions such as negation, e.g., "*Elephants are not small*" is a GENERIC SENTENCE. SE types aim to capture how clauses behave in discourse, and the types STATE and EVENT are aspectual rather than ontological categories. The types reflect not so much semantic content of a clause as its manner of presentation, and all parts of a clause contribute to determining its SE type.

[1] Corpora, annotation manual and code available at www.coli.uni-saarland.de/projects/sitent

3 Related work

SE types model aspect at the clause level; thus they are most closely related to other works performing automatic classification for various aspect-related phenomena of the verb or the clause. For example, Vendler classes (Vendler, 1957) ascribe four categories as lexical properties of verbs, distinguishing states from three types of events (accomplishment, achievement, and activity), differing according to temporal and aspectual properties (e.g. telicity and punctuality). The work of Siegel and McKeown (2000) is a major inspiration in computational work on modeling these linguistic phenomena, introducing the use of linguistic indicators (see Section 5.1). Hermes et al. (2015) model Vendler classes computationally on a verb-type level for 95 different German verbs, combining distributional vectors with supervised classification. Zarcone and Lenci (2008) investigate both supervised and unsupervised classification frameworks for occurrences of 28 Italian verbs, and Friedrich and Palmer (2014a) predict lexical aspectual class for English verbs in context.

The only previous approach to automatic classification of SE types comes from Palmer et al. (2007). This system (UT07) uses word and POS tag features as well as a number of lexical features adopted from theoretical work on aspectual classification. The model is described in Section 6.1.

Another related body of work has to do with determining event class as a precursor to temporal relation classification. The inventory of event classes, described in detail in the TimeML annotation guidelines (Saurí et al., 2006), combines semantic (REPORTING, PERCEPTION), aspectual (ASPECTUAL, STATE, OCCURRENCE), and intensional (I_ACTION, I_STATE) properties of events.

Finally, there are close connections to systems which predict genericity of noun phrases (Reiter and Frank, 2010; Friedrich and Pinkal, 2015a), and habituality of clauses (Mathew and Katz, 2009; Friedrich and Pinkal, 2015b).

4 Data sets

The experiments presented in this paper make use of two data sets labeled with SE types.

Brown data. This data set consists of 20 texts from the *popular lore* section of the Brown corpus (Francis and Kučera, 1979), manually segmented into 4391 clauses and marked by two annotators in an intuitive way with κ=0.52 (Cohen, 1960). Final labels were created via adjudication. The texts are essays and personal stories with topics ranging from maritime stories to marriage advice.

corpus	tokens	SEs	Fleiss' κ
MASC	357078	30333	0.69
Wikipedia	148040	10607	0.66

Table 1: SE-labeled corpora: **size** and **agreement**.

SE type	MASC	Wiki	Fleiss κ*
STATE	49.8	24.3	0.67
EVENT	24.3	18.9	0.74
REPORT	4.8	0.9	0.80
GENERIC	7.3	49.7	0.68
GENERALIZING	3.8	2.5	0.43
QUESTION	3.3	0.1	0.91
IMPERATIVE	3.2	0.2	0.94
undecided	2.4	2.1	-

Table 2: **Distribution of SE types** in gold standard (%). *Krippendorff's diagnostics.

MASC and Wikipedia. Our main data set consists of documents from MASC (Ide et al., 2010) and Wikipedia. The MASC data covers 12 of the written genres (see Table 11). Texts are split into clauses using SPADE (Soricut and Marcu, 2003) with some heuristic post-processing, and the clauses are labeled by three annotators independently. Annotators, all student assistants with basic linguistic training, were given an extensive annotation manual. Table 1 reports agreement over types in terms of Fleiss' κ (Fleiss, 1971). As we do not force annotators to assign a label to each clause, we compute κ using all pairings of labels where both annotators assigned an SE type. The gold standard is constructed via majority voting.

Table 2 shows the distribution of SE types. The largest proportion of segments in MASC are STATEs, while the largest proportion in Wikipedia are GENERIC SENTENCEs. The Wikipedia data was collected to supplement MASC, which contains few generics and no data from an encyclopedic genre. Within MASC, the various genres' distributions of SE types differ as well, and agreement scores also vary: some genres contain many instances of easily classifiable SE types, while others (e.g., essays or journal) are more difficult to annotate (more details in Section 6.6).

The rightmost column of Table 2 shows the values for Krippendorff's diagnostics (Krippendorff, 1980), a tool for determining which categories hu-

group	explanation	examples
mv	features describing the SE's **main verb** & its arguments	tense, lemma, lemma of object, auxiliary, WordNet sense and hypernym sense, progressive, POS, perfect, particle, voice, linguistic indicators
mr	features describing the **main referent**, i.e., the NP denoting the main verb's subject	lemma, determiner type, noun type, number, WordNet sense and supersense, dependency relations linked to this token, person, countability, bare plural
cl	features describing entire **clause** that invokes the SE	presence of adverbs / prepositional clauses, conditional, modal, whether subject before verb, negated, verbs embedding the clause

Table 3: Overview of feature set **B**. The full and detailed list is available (together with the implementation) at `http://www.coli.uni-saarland.de/projects/sitent`.

mans had most difficulties with. For each category, one computes κ for an artificial set-up in which all categories except one are collapsed into an artificial OTHER category. A high value indicates that annotators can distinguish this SE type well from others. GENERALIZING SENTENCEs are most difficult to agree upon. For all frequently occurring types as well as QUESTIONs and IMPERATIVEs, agreement is substantial.

Agreement on QUESTION and IMPERATIVEs is not perfect even for humans, as identifying them requires recognizing cases in reported speech, which is not a trivial task (e.g., Brunner, 2013). To illustrate another difficult case, consider the example "You must never confuse faith", which was marked as both IMPERATIVE and GENERIC SENTENCE, by different annotators.

5 Method

This section describes the feature sets and classification methods used in our approach, which models SE type labeling as a supervised sequence labeling task.

5.1 Feature sets

Our feature sets are designed to work well on large data sets, across genres and domains. Features are grouped into two sets: **A** consists of standard NLP features including POS tags and Brown clusters. Set **B** targets SE labeling, focusing on syntactic-semantic properties of the main verb and main referent, as well as properties of the clause which indicate its aspectual nature. Texts are pre-processed with Stanford CoreNLP (Manning et al., 2014), including tokenization, POS tagging (Toutanova et al., 2003) and dependency parsing (Klein and Manning, 2002) using the UIMA-based DKPro framework (Ferrucci and Lally, 2004; Eckart de Castilho and Gurevych, 2014).

A-pos: part-of-speech tags. These features count how often each POS tag occurs in a clause.

A-bc: Brown cluster features. UT07 relies mostly on words and word/POS tag pairs. These simple features work well on the small Brown data set, but the approach quickly becomes impractical with increasing corpus size. We instead turn to distributional information in the form of Brown clusters (Brown et al., 1992), which can be learned from raw text and represent word classes in a hierarchical way. Originally developed in the context of n-gram language modeling, they aim to assign words to classes such that the average mutual information of the words in the clusters is maximized. We use existing, freely-available clusters trained on news data by Turian et al. (2010) using the implementation by Liang (2005).[2] Clusterings with 320 and 1000 Brown clusters work best for our task. We use one feature per cluster, counting how often a word in the clause was assigned to this cluster (0 for most clusters).

B-mv: main verb. Using dependency parses, we extract the verb ranked highest in the clause's parse as the main verb, and extract the set of features listed in Table 3 for that token. Features based on WordNet (Fellbaum, 1998) use the most frequent sense of the lemma. Tense and voice information is extracted from sequences of POS tags using a set of rules (Loaiciga et al., 2014). Linguistic indicators (Siegel and McKeown, 2000) are features collected per verb type over a large parsed background corpus, encoding how often a verb type occurred with each linguistic marker, e.g., in past tense or with an *in*-PP. We use values collected from Gigaword (Graff et al., 2003); these are freely available at our project web site (Friedrich and Palmer, 2014a).

B-mr: main referent. We extract the grammatical subject of the main verb (i.e., `nsubj` or `nsubjpass`) as the clause's main referent. While the main verb must occur within the clause, the

[2] `http://metaoptimize.com/projects/wordreprs`

main referent may be a token either within or outside the clause. In the latter case, it still functions as the clause's main referent, as in most cases it can be considered an implicit argument within the clause. Table 3 lists the features extracted for the main referent.

B-cl: clause. These features (see also Table 3) describe properties at the clause level, capturing both grammatical phenomena such as word order and lexical phenomena including presence of particular adverbials or prepositional phrases, as well as semantic information such as modality. If the clause's main verb is embedded in a `ccomp` relation, we also use features describing the respective governing verb.

5.2 Classification / sequence labeling model

Our core modeling assumption is to view a document as a sequence of SE type labels, each associated with a clause; this motivates the choice of using a conditional random field (CRF, Lafferty et al. (2001)) for label prediction. The conditional probability of label sequence $\vec{y}$ given an observation sequence $\vec{x}$ is given by:

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} exp(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)),$$

with $Z(\vec{x})$ being a normalization constant (see also Klinger and Tomanek (2007)). λ_i, the weights of the feature functions, are learned via L-BGFS (Wright and Nocedal, 1999).

We create a linear chain CRF using the CRF++ toolkit[3] with default parameters, applying two forms of feature functions: $f_i(y_j, x_j)$ and $f_i(y_{j-1}, y_j)$. The former consists of indicator functions for combinations of SE type labels and each of the features listed above. The latter type of feature function (also called "bigram" features in CRF++ terminology) gets instantiated as indicator functions for each combination of labels, thereby enabling the model to take sequence information into account. When using only the former type of feature function, our classifier is equivalent to a maximum entropy (MaxEnt) model.

Side remark: pipeline approach. Feature set **B** is inspired by previous work on two subtasks of assigning an SE type to a clause (see Section 3): (a) identifying the genericity of a noun phrase in its clausal context, and (b) identifying whether a clause is episodic, habitual or static. This information can in turn be used to determine a clause's SE type label in a rule-based way, e.g., GENERALIZING SENTENCEs are habitual clauses with a non-generic main referent. As our corpus is also annotated with this information, we also trained separate models for these subtasks and assigned the SE type label accordingly. However, such a pipeline approach is not competitive with the model trained directly on SE types (see Section 6.3).

[3] `https://code.google.com/p/crfpp`

6 Experiments

Here we present our experiments on SE type classification, beginning with a (near) replication of the UT07 system, and moving on to evaluate our new approach from multiple perspectives.

6.1 Replication and extension of UT07

As a first step, we implement a system similar to UT07, which relies on the features summarized in Table 4. For W and T features, we set a frequency threshold of 7 occurrences. Feature set L comprises sets of predicates assumed to correlate with particular SE types, and whether or not the clause contains a modal or finite verb. Set G includes all verbs of the clause and their POS-tags. UT07 additionally uses CCG supertags and grammatical function information. The UT07 system approximates a sequence labeling model by adding the predicted labels of previous clauses as *lookback* (LB) features. To parallel their experiments, we train both MaxEnt and CRF models, as explained in Section 5.2. Results on the Brown data, with the same training/test split, appear in Table 5. Unlike the LB model, our CRF predicts the label sequence jointly and outperforms UT07 on the Brown data by up to 7% accuracy. We assume that the performance boost in the MaxEnt setting is at least partially due to having better parses.

In sum, on the small Brown data set, a CRF approach successfully leverages sequence information, and a simple set of features works well. Preliminary experiments applying our new features on Brown data yield no improvements, suggesting that word-based features overfit this domain.

W	words
T	POS tags, word/POS tag combinations
L	linguistic cues
G	grammatical cues

Table 4: Features used in baseline UT07.

features	Palmer et al. (2007) MaxEnt	LB	our implementation MaxEnt	CRF
W	45.4	46.6	48.8	47.0
WT	49.9	52.4	52.9	53.7
WTL	48.9	50.5	51.6	55.8
WTLG	**50.6**	**53.1**	**55.8**	**60.0**

Table 5: **Accuracy** on Brown. Test set majority class (STATE) is 35.3%. LB = results for best look-back settings in MaxEnt. 787 test instances.

6.2 Experimental settings

We develop our models using **10-fold cross validation (CV)** on 80% (counted in terms of the number of SEs) of the MASC and Wikipedia data (a total of 32,855 annotated SEs), keeping the remaining 20% as a held-out test set. Development and test sets each contain distinct sets of documents; the documents of each MASC genre and of Wikipedia are distributed over the folds. We report results in terms of macro-average precision, macro-average recall, macro-average F1-measure (harmonic mean of macro-average precision and macro-average recall), and accuracy. We apply McNemar's test (McNemar, 1947) with $p < 0.01$ to test significance of differences in accuracy. In the following tables, we mark numerically-close scores with the same symbols if they are found to be significantly different.

Upper bound: human performance. Labeling clauses with their SE types is a non-trivial task even for humans, as there are many borderline cases (see Sections 4 and 8). We compute an upper bound for system performance by iterating over all clauses: for each pair of human annotators, two entries are added to a co-occurrence matrix (similar to a confusion matrix), with each label serving once as "gold standard" and once as the "prediction." From this matrix, we can compute scores in the same manner as for system predictions. Precision and recall scores are symmetric in this case, and accuracy corresponds to observed agreement.

6.3 Impact of feature sets

We now compare various configurations of our CRF-based SE type labeler, experimenting with the feature sets as described in Section 5.1. Table 6 shows the results for 10-fold CV on the dev portion of the MASC+Wiki corpus.

Each feature set on its own outperforms the majority class baseline. Of the individual feature groups, **bc** and **mv** have the highest predictive power; both capture lexical information of the main verb. Using sets **A** and **B** individually results in similar scores; their combination increases accuracy on the dev set by an absolute 3.6-4.3%. Within **A** and **B**, each subgroup contributes to the increase in performance (not shown in table).

feature set	P	R	F	acc.
maj. class (STATE)	6.4	14.3	8.8	45.0
A	70.1	61.4	65.4	∗†72.1
pos	49.3	40.3	44.3	58.7
bc	67.5	55.8	61.1	∗70.6
B	69.5	62.7	66.9	⋆‡72.8
mr	36.4	26.8	30.9	51.7
mv	62.3	52.4	56.9	⋆70.8
cl	53.3	41.2	46.6	52.8
A+B	**74.1**	**68.6**	**71.2**	‡†**76.4**
upper bound (humans)	78.6	78.6	78.6	79.6

Table 6: **Impact of different feature sets.** Wiki+MASC dev set, CRF, 10-fold CV.

feature set	P	R	F	acc.
maj. class (STATE)	6.4	14.3	8.8	44.7
A	67.6	60.6	63.9	∗69.8
B	69.9	61.7	65.5	†71.4
A+B	**73.4**	**65.5**	**69.3**	∗†**74.7**

Table 7: Results on MASC+Wiki **held-out test set** (7937 test instances).

Finally, having developed exclusively on the dev set, we run the system on the held-out test set, training on the entire dev set. Results (in Table 7) show the same tendencies as for the dev set: each feature set contributes to the final score, and the syntactic-semantic features targeted at classifying SE types (i.e. **B**) are helpful.

Ablation. To gain further insight, we ablate each feature subgroup from the full system, see Table 8. Again, **bc** features and **mv** features are identified as the most important ones. The other feature groups partially carry redundant information when combining **A** and **B**. Next, we rank features by their information gain with respect to the SE types. In Table 3, the features of each group are ordered by this analysis. Ablating single features from the full system does not result in significant performance losses. However, using only selected, top features for our system decreased performance, possibly because some features cover rare but important cases, and because the feature selection algorithm does not take into account the information features may provide regarding tran-

feature set	P	R	F	acc.
A+B	**74.1**	**68.6**	**71.2**	**76.4**
- bc	71.3	65.7	68.4	74.5
- pos	73.4	67.4	70.2	76.0
- mr	73.7	67.4	70.4	75.9
- mv	72.3	64.2	68.0	73.6
- cl	73.1	67.6	70.2	76.0

Table 8: **Impact of feature groups: ablation** Wiki+MASC dev set, CRF, 10-fold CV. All accuracies for ablation settings are significantly different from A+B.

sitions (Klinger and Friedrich, 2009). In addition, CRFs are known to be able to deal with a large number of potentially dependent features.

Side remark: pipeline approach. We here use the subset of SEs labeled as EVENT, STATE, GENERIC SENTENCE or GENERALIZING SENTENCE because noun phrase genericity and habituality is not labeled for IMPERATIVE and QUESTION, and REPORT is identified lexically based on the main verb rather than these semantic features. Models for subtasks of SE type identification, i.e., (a) genericity of noun phrases and (b) habituality reach accuracies of (a) 86.8% and (b) 83.6% (on the relevant subset). Applying the labels output by these two systems as (the only) features in a rule-based way using a J48 decision tree (Witten et al., 1999) results in an accuracy of 75.5%, which is lower than 77.2%, the accuracy of the CRF which directly models SE types (when using only the above four types).

6.4 Impact of amount of training data

Next we test how much training data is required to get stable results for SE type classification. Figure 3 shows accuracy and F1 for 10-fold CV using **A+B**, with training data downsampled to different extents in each run by randomly removing documents. Up to the setting which uses about 60% of the training data, performance increases steadily. Afterwards, the curves start to level off. We conclude that robust models can be learned from our corpus. Adding training data, especially in-domain data, will, as always, be beneficial.

6.5 Impact of sequence labeling approach

Palmer et al. (2007) suggest that SE types of nearby clauses are a useful source of information. We further test this hypothesis by comparing our sequence labeling model (CRF) to two additional

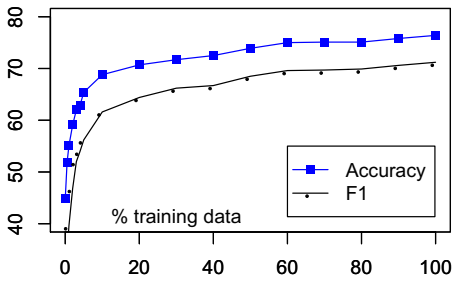

Figure 3: **Learning curve** for MASC+Wiki dev.

models: (1) a MaxEnt model, which labels clauses in isolation, and (2) a MaxEnt model including the correct label of the preceding clause (seq-oracle), simulating an upper bound for the impact of sequence information on our system.

Table 9 shows the results. Scores for GENERALIZING SENTENCE are the lowest as this class is very infrequent in the data set. The most striking improvement of the two sequence labeling settings over MaxEnt concerns the identification of GENERIC SENTENCES. These often "cluster" in texts (Friedrich and Pinkal, 2015b) and hence their identification profits from using sequence information. The results for seq-oracle show that the sequence information is useful for STATE, GENERIC and GENERALIZING SENTENCES, but that no further improvement is to be expected from this method for the other SE types. We conclude that the CRF model is to be preferred over the MaxEnt model; in almost all of our experiments it performs significantly better or equally well.

SE type	MaxEnt	CRF	seq-oracle
STATE	79.1	**80.6**	81.7
EVENT	77.5	**78.6**	78.3
REPORT	78.2	**78.9**	78.3
GENERIC	61.3	**68.3**	73.5
GENERALIZING	25.0	**29.4**	38.1
IMPERATIVE	72.3	**75.3**	74.7
QUESTION	84.4	**84.4**	83.8
macro-avg P	71.5	**74.1**	75.5
macro-avg R	66.1	**68.6**	70.4
macro-avg F1	68.7	**71.2**	73.9
accuracy	∗74.1	∗†**76.4**	†77.9

Table 9: **Impact of sequence information:** (**F1** by SE type): CRF, Masc+Wiki, 10-fold CV.

6.6 Impact of genre

In this section, we test to what extent our models are robust across genres. Table 10 shows F1-scores for each SE type for two settings: the 10-

SE type	genre-CV	10-fold CV	humans
STATE	78.2	80.6	82.8
EVENT	77.0	78.6	80.5
REPORT	76.8	78.9	81.5
GENERIC	44.8	68.3	75.1
GENERALIZING	27.4	29.4	45.8
IMPERATIVE	70.8	75.3	93.6
QUESTION	81.8	84.4	90.7
macro-avg F1	66.6	71.2	78.6
accuracy	∗71.8	∗76.4	79.6

Table 10: **Impact of in-genre training data**. F1-score by SE type, CRF, MASC+Wiki dev.

fold CV setting as explained in section 6.2, and a **genre-CV** setting, simulating the case where no in-genre training data is available, treating each genre as one cross validation fold. As expected, in the latter setting, both overall accuracy and macro-average F1 are lower compared to the case when in-genre training data is available. Nevertheless, our model is able to capture the nature of SE types across genres: the prediction of STATE, EVENT, REPORT and QUESTION is relatively stable even in the case of not having in-genre training data. An EVENT seems to be easily identifiable regardless of the genre. GENERIC SENTENCE is a problematic case; in the genre-CV setting, its F1-score drops by 23.5%. The main reason for this is that the distribution of SE types in Wikipedia differs completely from the other genres (see section 4). Precision for GENERIC SENTENCE is at 70.5% in the genre-CV setting, but recall is only 32.8% (compared to 70.1% and 66.6% in the 10-fold CV setting). Genericity seems to work differently in the various genres: most generics in Wikipedia clearly refer to kinds (e.g., lions or plants), while many generics in essays or letters are instances of more abstract concepts or generic *you*.

Results by genre. Next, we drill down in the evaluation of our system by separately inspecting results for individual genres. Table 11 shows that performance differs greatly depending on the genre. In some genres, the nature of SE types seems clearer to our annotators than in others, and this is reflected in the system's performance. The majority class is GENERIC SENTENCE in wiki, and STATE in all other genres. In the 'same genre' setting, 10-fold CV was performed within each genre. Adding out-of-genre training data improves macro-average F1 especially for genres with low scores in the 'same genre' setting. This boost is

		training data			
	maj. class	same genre	all	humans	
genre	%	F1	F1	F1	κ
blog	57.6	57.3	**64.9**	72.9	0.62
email	68.6	63.6	**66.4**	67.0	0.65
essays	49.4	33.5	**62.1**	64.6	0.54
ficlets	44.7	60.2	**65.7**	81.7	0.80
fiction	45.8	63.0	**66.0**	76.7	0.77
govt-docs	60.9	26.6	**67.6**	72.6	0.57
jokes	34.9	66.2	**69.8**	82.0	0.77
journal	59.3	35.8	**59.8**	63.7	0.52
letters	57.3	51.9	**65.1**	68.0	0.66
news	52.2	54.6	**64.1**	78.6	0.75
technical	57.7	31.4	**59.4**	54.7	0.55
travel	25.9	39.9	**58.1**	48.9	0.59
wiki	51.6	53.1	**63.0**	69.2	0.66

Table 11: **Macro-avg. F1 by genre**, CRF, 10-fold CV. Majority class given in % of clauses.

due to adding training data for types that are infrequent in that genre. Accuracy (not shown in table) improves significantly for blog, essays, govt-docs, jokes, and journal, and does not change for the remaining genres. We conclude that it is extremely beneficial to use our full corpus for training, as robustness of the system is increased, especially for SE types occurring infrequently in some genres.

7 Identification of Abstract Entities

Our system notably does not address one of Smith's main SE categories: *Abstract Entities*, introduced in Section 2. These SEs are expressed as clausal arguments of certain predicates such as (canonically) *know* or *believe*. Note that the *Abstract Entity* subtypes FACT and PROPOSITION do not imply that a clause's propositional content is true or likely from an objective point of view, they rather indicate that the clause's content is introduced to the discourse as an object of knowledge or belief, respectively. Following Smith (2003), we use PROPOSITION in a different sense than the usual meaning of "proposition" in semantics - naturally situation entities of any type may have propositional content. Smith's use of the term (and thus ours too) contrasts PROPOSITION with FACT - our PROPOSITIONS are simply sentences presented as a belief (or with uncertain evidential status) of the writer or speaker. This use of "proposition" also occurs in linguistic work by Peterson (1997) on factive vs. propositional predicates.

During the creation of the corpus, annotators

were asked to give one label out of the SE types included in our classification task, and to mark the clause with one of the *Abstract Entity* subtypes in addition if applicable. Analysis of the data shows that our annotators frequently forgot to mark clauses as *Abstract Entities*, which makes it difficult to model these categories correctly. As a first step toward resolving this issue, we implement a filter which automatically identifies *Abstract Entities* by looking for clausal complements of certain predicates. The list of predicates is compiled using WordNet synonyms of *know*, *think*, and *believe*, as well as predicates extracted from FactBank (Sauri and Pustejovsky, 2009) and TruthTeller (Lotan et al., 2013). Many of the clauses automatically identified as *Abstract Entities* are cases that annotators missed during annotation. We thus performed a post-hoc evaluation, presenting these clauses in context to annotators and asking whether the clause is an *Abstract Entity*. The so-estimated precision of our filter is 85.8% (averaged over 3 annotators). Agreement for this annotation task is $\kappa = 0.54$, with an observed agreement of 88.7%. Our filter finds 80% of the clauses labeled as *Abstract Entity* by at least one annotator in the gold standard; this is *approximately* its recall.

8 Conclusion

We have presented a system for automatically labeling clauses with their SE type which is mostly robust to changes in genre and which reaches accuracies of up to 76%, comparing favorably to the human upper bound of 80%. The system benefits from capturing contexual effects by using a linear chain CRF with label bigram features. In addition, the distributional and targeted syntactic-semantic features we introduce enable SE type prediction for large and diverse data sets. Our publicly available system can readily be applied to any written English text, making it easy to explore the utility of SE types for other NLP tasks.

Discussion. Our annotation scheme and guidelines for SE types (Friedrich and Palmer, 2014b) follow established traditions in linguistics and semantic theory. When applying these to a large number of natural texts, though, we came across a number of borderline cases where it is not easy to select just one SE type label. As we have reported before (Friedrich et al., 2015), the most difficult case is the identification of GENERIC SENTENCEs, which are defined as making a statement about a kind or class. We find that making this task becomes particularly difficult for argumentative essays (Becker et al., to appear).

Future work. A next major step in our research agenda is to integrate SE type information into various applications, including argument mining, temporal reasoning, and summarization. Together with the mode of progression through the text, e.g., temporal or spatial, SE type distribution is a key factor for a reader or listener's intuitive recognition of the discourse mode of a text passage. Therefore the automatic labeling of clauses with their SE type is a prerequisite for automatically identifying a passage's discourse mode, which in turn has promising applications in many areas of NLP, as the mode of a text passage has implications for the linguistic phenomena to be found in the passage. Examples include temporal processing of text (Smith, 2008), summarization, or machine translation (for work on genres see van der Wees et al., 2015). Here we focus on the automatic identification of SE types, leaving the identification of discourse modes to future work.

The present work, using the SE type inventory introduced by Smith (2003), is also the basis for research on more fine-grained aspectual type inventories. Among others, we plan to create subtypes of the STATE label, which currently subsumes clauses stativized by negation, modals, lexical information or other aspectual operators. Distinguishing these is relevant for temporal relation processing or veridicality recognition.

Acknowledgments

We thank the anonymous reviewers and Andrea Horbach for their helpful comments related to this work, and our annotators Melissa Peate Sørensen, Christine Bocionek, Kleo-Isidora Mavridou, Fernando Ardente, Damyana Gateva, Ruth Kühn and Ambika Kirkland. This research was supported in part by the MMCI Cluster of Excellence of the DFG, and the first author is supported by an IBM PhD Fellowship. The second author is funded by the Leibniz ScienceCampus *Empirical Linguistics and 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

Emmon Bach. 1986. The algebra of events. *Linguistics and philosophy* 9(1):5–16.

Maria Becker, Alexis Palmer, and Anette Frank. to appear. Argumentative texts and clause types. In *Proceedings of the 3rd Workshop on Argument Mining, ACL 2016*.

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):467–479.

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and linguistic computing* 28(4):563–575.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Dublin, Ireland.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 10(3-4):327–348.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.

W. Nelson Francis and Henry Kučera. 1979. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Brown University.

Annemarie Friedrich and Alexis Palmer. 2014a. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, USA.

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

Annemarie Friedrich, Alexis Palmer, Melissa Peate Srensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop (LAW IX)*. Denver, Colorado, US.

Annemarie Friedrich and Manfred Pinkal. 2015a. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal.

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

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia* .

Jürgen Hermes, Michael Richter, and Claes Neuefind. 2015. Automatic induction of German aspectual verb classes in a distributional framework. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Charles Fillmore. 2008. MASC: The manually annotated sub-corpus of American English .

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 ACL 2010 conference short papers*.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*.

Roman Klinger and Christoph M Friedrich. 2009. Feature subset selection in conditional random fields for named entity recognition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

Roman Klinger and Katrin Tomanek. 2007. Classical probabilistic models and conditional random fields. *TU Dortmund Algorithm Engineering Report* .

Manfred Krifka, Francis Jeffrey Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link, and Gennaro Chierchia. 1995. Genericity: An Introduction. *The Generic Book* pages 1–124.

Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, University of California Berkeley.

Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *The Ninth Language Resources and Evaluation Conference (LREC)*.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. TruthTeller: Annotating Predicate Truth. In *Proceedings of NAACL 2013*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*.

Thomas A. Mathew and Graham E. Katz. 2009. Supervised categorization for habitual versus episodic sentences. In *Sixth Midwest Computational Lingustics Colloquium*. Indiana University, Bloomington, Indiana.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. 2007. A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Philip L Peterson. 1997. On representing event reference. In *Fact Proposition Event*, Springer, pages 65–90.

Nils Reiter and Anette Frank. 2010. Identifying Generic Noun Phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden.

Roser Saurí, Jessica Littman, Bob Knippen, Rober Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML Annotation Guidelines, Version 1.2.1. Technical report.

Roser Sauri and James Pustejovsky. 2009. Factbank 1.0 ldc2009t23. Web Download. Philadelphia: Linguistic Data Consortium.

Eric V Siegel and Kathleen R McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics* 26(4):595–628.

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

Carlota S Smith. 2005. Aspectual entities and tense in discourse. In *Aspectual inquiries*, Springer, pages 223–237.

Carlota S Smith. 2008. Time with and without tense. In *Time and modality*, Springer, pages 227–249.

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*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. 2015. What's

in a Domain? Analyzing Genre and Topic differences in Statistical Machine Translation. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics (ACL)*.

Benjamin D Van Durme. 2010. *Extracting implicit knowledge from text*. Ph.D. thesis, University of Rochester.

Zeno Vendler. 1957. Verbs and times. *The philosophical review* pages 143–160.

Ian H Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. Weka: Practical machine learning tools and techniques with java implementations .

Stephen Wright and Jorge Nocedal. 1999. Numerical optimization. *Springer Science* 35:67–68.

Alessandra Zarcone and Alessandro Lenci. 2008. Computational models of event type classification in context. In *Proceedings of LREC2008*.