Discourse-sensitive Automatic Identification of Generic Expressions

Annemarie Friedrich **Manfred Pinkal**
Department of Computational Linguistics
Saarland University, Saarbrücken, Germany
{afried,pinkal}@coli.uni-saarland.de

Abstract

This paper describes a novel sequence labeling method for identifying generic expressions, which refer to kinds or arbitrary members of a class, in discourse context. The automatic recognition of such expressions is important for any natural language processing task that requires text understanding. Prior work has focused on identifying generic noun phrases; we present a new corpus in which not only subjects but also clauses are annotated for genericity according to an annotation scheme motivated by semantic theory. Our context-aware approach for automatically identifying generic expressions uses conditional random fields and outperforms previous work based on local decisions when evaluated on this corpus and on related data sets (ACE-2 and ACE-2005).

1 Introduction

Distinguishing between statements about particular individuals or situations and generic sentences is an important part of human language understanding. Consider example (1): sentence (a) names characteristic attributes of a kind, which are inherent to every (typical) individual, and sentence (b) describes a specific individual.

(1) *(a) The modern domestic horse has a life expectancy of 25 to 30 years.* ***(generic)***

(a) Old Billy lived to the age of 62. ***(non-generic)***

The above example illustrates that generic and non-generic sentences differ substantially in their semantic impact and entailment properties. It can be inferred from sentence (1a) that a *typical* horse has a life expectancy of 25 to 30 years, and if we know that *Nelly* is a horse, we can infer that its life expectancy is 25 to 30 years. Sentence (1b) has no such properties, it only allows inferences about the particular individual *Old Billy*.

An automatic classifier that recognizes generic expressions would be extremely valuable for various kinds of natural language processing systems: for text understanding and question answering systems, through the improvement of textual entailment methods, and for systems acquiring machine-readable knowledge from text. Machine-readable knowledge bases have different representations for statements corresponding to generic knowledge about kinds and knowledge about specific individuals. The non-generic sentence (1b) roughly speaking provides ABox content for a machine-readable knowledge base, i.e., knowledge about particular instances, e.g, *"A is an instance of B / has property X"*. In contrast, the generic sentence (1a) feeds the TBox, i.e., knowledge of the form *"All B are C / have property X"*. Reiter and Frank (2010) provide a detailed discussion of the relevance of the distinction between classes and instances for automatic ontology construction.

In this paper, we present a new corpus annotated in a linguistically motivated way for genericity, and a context-sensitive computational model for labeling sequences of clauses or noun phrases (NPs) with their genericity status. Both manual annotation and automatic recognition of generic expressions are challenging tasks: virtually all NP types – definites, indefinites and quantified NPs, full NPs, pronouns, and even proper names (e.g. species names such as *Elephas maximus*) – can be found in generic and non-generic uses depending on their clausal context.

In this work, we call clauses generic if they provide a general characterization of entities of a certain kind, and we call mentions of NPs generic if they refer to kinds or arbitrary members of a class. Although genericity on the clause- and NP-level

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1272–1281,
Beijing, China, July 26-31, 2015.

are strongly interrelated, the concepts do not always coincide. As example (2) shows, sentences describing episodic events can have a generic NP as their subject. Note that references to species are kind-referring / generic on the NP level (following Krifka et al. (1995), see p. 65).

(2) *In September 2013 the blobfish was voted the "World's Ugliest Animal". (subject* ***generic****, clause* ***non-generic****)*

Genericity often cannot be annotated without paying attention to the wider discourse context. Clearly, coreference information is needed for the genericity classification of pronouns. Often, even genericity of full NPs or entire clauses cannot be decided in isolation, as illustrated by example (3). Sentence (b) could be part of a particular narrative about a tree, or it could be a generic statement. Only the context given by (a) clarifies that (b) indeed makes reference to any year's new twigs and is to be interpreted as generic.

(3) *(a) Sugar maples also have a tendency to color unevenly in fall. (****generic****)*
*(b) The recent year's growth twigs are green and turn dark brown. (****generic****)*

In computational linguistics, most research on detecting genericity has been done in relation to the ACE corpora (Mitchell et al., 2003; Walker et al., 2006), focusing on assigning genericity labels to noun phrases (Suh et al., 2006; Reiter and Frank, 2010), see Section 2. Our work is based on these approaches, most notably on the work of Reiter and Frank (2010), and extends upon them in the following essential ways.

The major contributions of this work are: (1) We create *a new corpus of Wikipedia articles* annotated with linguistically motivated genericity labels both on the subject- and clause-level (see Section 3). The corpus is balanced with respect to genericity and about 10,000 clauses in size. (2) We present a *discourse-sensitive genericity labeler*. Technically, we use conditional random fields as a sequence labeling method (Section 4). We train and evaluate our method on the Wikipedia dataset and the ACE corpora, evaluating both the tasks of predicting NP genericity and the task of predicting clause-level genericity. Our labeler outperforms the state-of-the-art by a margin of 6.6-11.9% (depending on the data set) in terms of accuracy, at the same time increasing F_1-score. Much of the performance gain is due to the inclusion of discourse information. For the discussion of our experimental results, see Section 5.

In this paper, we do *not* address the following two important aspects of genericity. First, *habitual* sentences form a class of generalizing statements which bear a close relation to generics. As can be seen in example (4), they describe a characterizing property of either a specific entity or a class by generalizing over situations instead of or in addition to entities (Carlson, 2005). We classify habitual sentences with a generic subject as generic, and habitual sentences which describe a specific entity as non-generic, leaving the task of habituality detection for future work.

(4) *(a) John smokes after dinner.*
(b) Gentlemen smoke after dinner.

Second, generic clauses express regularities within classes of entities, and thus are similar to universally quantified sentences in their truth conditions and entailment properties. However, their truth-conditional interpretation is tricky, since they express typicality, describe stereotypes and allow exceptions, for example *Dutchmen are good sailors* is not false even if most Dutchmen do not sail at all (Carlson, 1977). We concentrate on the decision of whether a clause is generic or not, and leave the truth-conditional interpretation for further work. For a detailed discussion of the semantics of generics expressions see the comprehensive survey by Krifka et al. (1995); a short and instructive overview can be found in the first part of (Reiter and Frank, 2010).

2 Related Work

In this section, we first briefly review previously developed annotation schemes for genericity. We then describe work on automatically predicting the genericity of NPs or different types of clauses.

Annotation. ACE-2 (Mitchell et al., 2003) and ACE-2005 (Walker et al., 2006) are the two most notable annotation projects for labeling genericity of NPs to date. In the ACE-2 corpus, 40106 entity mentions in 520 newswire and broadcast documents are marked with regard to whether they refer to "any member of the set in question" (GEN, generic) rather than "some particular, identifiable member of that set" (SPC, specific/non-generic). The major drawback of ACE-2 is that genericity is basically defined as lack of specificity, which leads

to uncertainty and inconsistencies in the annotation process, and to a heterogeneous set of NPs labeled with GEN, including quantificational NPs and NPs in modalized, future, conditional, hypothetical, negated, uncertain, and question contexts. In addition, in both ACE-2 and ACE-2005, predicative and modifier uses of nouns, to which the genericity distinction is not applicable, also receive labels (e.g. *John seems to be a nice person / a subway system*).

In the updated guidelines of ACE-2005, the label USP (underspecified) is introduced for non-generic non-specific reference, including NPs in the various contexts mentioned above that were improperly labeled as generic in ACE-2. The class also contains mentions of an entity whose identity would be 'difficult to locate' (*Officials reported ...*). Moreover, annotators are asked to mark truly ambiguous cases that have both a generic and a non-generic reading as USP. Finally, NEG (negated) marks negatively quantified entities that refer to the empty set of the kind mentioned.

While we agree that in general there are underspecified cases, the guidelines for ACE-2005 mix other phenomena into the USP class, resulting in a high confusion between USP and both of the labels SPC and GEN in the manual annotations (Friedrich et al., 2015). Data from two annotators is available, and we compute an agreement of Cohen's $\kappa = 0.53$ over the four labels. The ACE corpora consist only of news data, and the distributions of labels are highly skewed towards specific mentions. For some criticism of the ACE annotation scheme, see also Suh (2006).

Several linguistically motivated annotation studies targeting genericity of noun phrases bear similarity to our annotation scheme (Section 3), but comprise very little data (Poesio, 2004; Herbelot and Copestake, 2009). In the ARRAU corpus (Poesio and Artstein, 2008), about 24321 markables are tagged for genericity.

Nedoluzhko (2013) survey the treatment of genericity phenomena within coreference resolution research; they find a consistent definition of genericity to be lacking. Friedrich and Palmer (2014b) present an annotation scheme for situation types including generic sentences, which they find to be infrequent in their corpus consisting of news, jokes and (fund-raising) letters. Our new WikiGenerics corpus contains more than 10,000 clauses, approximately half of which are generic.

Automatic Identification of Genericity. Suh et al. (2006) propose a rule-based approach, which extracts only bare plurals and singular NPs quantified with *every* or *any* as generic. Reiter and Frank (2010) use a wide range of syntactic and semantic features to train a supervised classifier for identifying generic NPs. We compare to their method (described in detail in Section 5.2) as a highly-competitive baseline.

Palmer et al. (2007) classify clauses into several types of situation entities including states, events, generalizing sentences (habitual utterances referring to specific individuals) and generic sentences. They find that using context by using the labels of preceding clauses as features improves the classification of clause types, but generic sentences are extremely sparse in their data set. Our present approach uses a sequence labeling model that computes the best labeling for an entire sequence.

3 WikiGenerics: Data and Annotations

In order to study generics in a genre other than news (as in ACE), we turn to an encyclopedia, in which we expect many generics. We create our **WikiGenerics** corpus[1] as follows. We aim to create a corpus that is balanced in the sense that it contains many generic and non-generic sentences, and also generics from many different domains. We collect 102 texts about animals, organised crime, ethnic groups, games, sports, medicine, music, politics, religion, scientific disciplines and biographies from Wikipedia. For example, some sentences make statements about a 'natural' kind (*Blobfish are typically shorter than 30 cm*), others express definitions such as the rules of a football game (*The offensive team must line up in a legal formation before they can snap the ball*).

Generic clauses have the typical form of a predicative statement about the sentence topic, which is normally realized as the grammatical subject in English. Intuitions about NP-level genericity and its relation to clause-level genericity are quite reliable for topic NPs of clauses, which also typically occur in subject position in English. Since generics in non-subject positions are less frequent and hard to interpret (see the discussion of "dependent generics" by Link (1995)), we decided to annotate subject NPs only. We are aware that we are missing relevant cases (e.g. the less preferred reading

[1]The WikiGenerics corpus is freely available at: www.coli.uni-saarland.de/projects/sitent

of *Cats chase mice*, which attributes to mice the property of being chased by cats), but in this work, we want to study the "easier" subject cases as a first step.

We use the discourse parser SPADE (Soricut and Marcu, 2003) to automatically segment the first 70 sentences of each article into clauses. Each clause is manually annotated with the following information (for more details on the annotation scheme, see (Friedrich et al., 2015)):

- **Task NP**: whether or not the *subject* NP of the clause refers to a class or kind (**generic** vs. **non-generic**);
- **Task Cl**: whether the *clause* is **generic**, defined as a clause that makes a characterizing statement about a class or kind, or **non-generic**.
- **Task Cl+NP**: using the information from Task NP and Cl above, we automatically derive the following classification for each *clause* (compare to the explanation of example (2)).
  - **GEN_gen**: generic clause, subject is generic by definition (*The lion is a predatory cat*);
  - **NON-GEN_non-gen**: non-generic clause with a non-generic subject (*Simba roared*);
  - or **NON-GEN_gen**: episodic clause with a generic subject (*Dinosaurs died out*).
  - **GEN_non-gen** does not exist by definition.

We construct the gold standard for our experiments via majority voting over the labels given by three paid annotators, students of computational linguistics. Annotators were given a written manual and a short training on documents not included in the corpus. They are given the option to indicate segmentation errors, e.g. that two segments should actually be one, or that one segment contains multiple clauses. In the latter case, we ask them to give labels for the first clause in the segment. 10240 (86%) of all pre-segmented clauses received labels for all three tasks from all annotators, who were allowed to skip clauses that do not contain a finite verb. Our gold standard includes an additional 115 segments that did not receive a label by one annotator but were unanimously labeled by the other two. The other segments are disregarded in the experiments. Some of them have expletive subjects, and most others are non-finite verb phrases such as *to*-infinites or headlines that consist of only a NP. Inter-annotator agreement measured as Fleiss' κ (Fleiss, 1971) on the segments labeled by all three annotators is 0.70, 0.73 and 0.69 for Task NP, Task Cl and Task Cl+NP respectively, indicating substantial agreement (Landis and Koch, 1977).

4 A Sequence Labeling Model for Genericity

This section describes our method for identifying generic clauses and NPs in context. We apply the following methods on each of the three different prediction tasks NP, Cl and Cl+NP introduced in Section 3, varying only the type of labels on which we train and test. In contrast to prior work, our computational model integrates not only information from each local instance, but also information about the genericity status of surrounding instances. The final labeling for the sequence of instances of an entire document is optimized with regard to these two types of information, which, as we have argued in Section 1, both play a crucial role in determining genericity. The sequences to be labeled contain all clauses or NPs of a document. We also tried labeling sequences for paragraphs instead of documents, but the performance was similar. A reason might be that paragraphs are quite often linked by mentioning the same entities (Friedrich and Palmer, 2014a).

Computational model. We use linear chain conditional random fields (Lafferty et al., 2001) to label sequences of mentions or sequences of clauses with regard to their genericity. Conditional random fields (CRFs) are well suited for our labeling task as they do not make an independence assumption between the features. CRFs predict the conditional probability of label sequence $\vec{y}$ given an observation sequence $\vec{x}$ as follows:

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

$Z(\vec{x})$ is a normalization constant, the sum over the scores of all possible label sequences for an observation sequence with the length of $\vec{x}$. The weights λ_i of the feature functions are the parameters to be learned. They do not depend on the current position j in the sequence. The feature functions f_i are in general allowed to look at the current label y_j, the previous label y_{j-1} and the entire observation sequence $\vec{x}$. We use a simple instantiation of a linear chain CRF whose feature functions take two forms, $f_i(y_j, x_j)$ and $f_i(y_{j-1}, y_j)$. We create a linear chain CRF model using the CRF++ toolkit[2], using all the default parameters.

[2] https://code.google.com/p/crfpp

NP-BASED FEATURES	
number	sg, pl
person	1, 2, 3
countability	from *Celex*, e.g. count
noun type	common, proper, pronoun
determiner type	def, indef, demon
part-of-speech	POS of head
bare plural	true, false
WN granularity	number of edges to top node
WN sense $[0-2]$	WN senses (head+hypernyms)
WN senseTop	top sense in hypernym hierarchy
WN lexical filename	person, artifact, event, ...
CLAUSE-BASED FEATURES	
dependency $[0-4]$	dependency relation between head and governor etc.
tense	tense, aspect and voice information, e.g. *pres_perf_active*
coarseTense	pres, past, fut
progressive	true, false
perfective	true, false
passive	true, false
temporal modifier	true, false
number of modifiers	numeric
part-of-speech	POS of head
predicate	lemma of head
adjunct-degree	positive, comparative, superlative
adjunct-pred	lemma of adverbial clauses' head

Table 1: **Features**. WN=WordNet.

Feature functions. We extract the set of features listed in Table 1 for each instance. This set of features is inspired by Reiter and Frank (2010), see also Section 5.2. In the case of the WikiGenerics corpus, the NP features are extracted for the subject of the clause. We parse the data using the Stanford parser (Klein and Manning, 2002) and obtain the subject NPs from the collapsed dependencies. For the ACE data, the NP features are extracted for all mentions in the gold standard and the clause features are extracted from the clause in which the mention appears. Our feature functions $f_i(y_j, x_j)$ are indicator functions combining the current label and one of the feature values of the current mention or clause, for example:

```
f = if (y_j = GENERIC and x_j.np.person=3)
      return 1 else return 0
```

We create two versions of the CRF model: the bigram[3] model additionally uses indicator functions $f(y_{j-1}, y_j)$ for each combination of labels, thus taking context into account. The unigram model does not use these feature functions, it is thus similar to a maximum entropy model (with a different normalization). Log-linear models work very well for many NLP tasks, especially if features are correlated as it is the case here, so in order to get a fair estimate of the impact of using the context (via the transition feature functions), we give numbers for this 'unigram' model in addition, rather than simply comparing the bigram-CRF to a Bayesian network, which is used by Reiter and Frank (2010). Using more complex feature functions did not result in significant performance gains, so we chose the simplest model. Note that even though the feature functions only formulate relationships between adjacent labels in the sequence, the optimal labeling is computed for the entire sequence: the choices of labels assigned to non-adjacent clauses *do* influence each other.

[3]Following CRF++ terminology.

Two-step Approach for Task Cl+NP. Task Cl+NP can be regarded as a combination of the two decisions made in Task NP and Task Cl. Therefore, we approach Task Cl+NP in two ways. (a) We train a CRF which directly outputs the three labels. (b) The *two-step* approach combines the output from the labelers trained for Task NP and Task Cl into one label in a rule-based way. This leads to the additional class **GEN_non-gen**, of which no gold instances exist by definition. As we evaluate in terms of F_1-score and accuracy for the existing classes, items classified into this artificial class will simply be counted as wrong and lack from the recall counts.

5 Experiments

This section reports on our experiments, which we evaluate in terms of precision (P), recall (R) and F_1-measure per class. We compute macro-averages as $P_{macro} = \frac{1}{|c|} * \sum_{i=1}^{|c|} P_i$ etc., where $|c|$ stands for the number of classes. Macro-F_1 is the harmonic mean of macro-average P and R. To report on statistical significance of differences in accuracy, we apply McNemar's test with $p < 0.01$.

5.1 Experimental Settings and Data

We report results for cross validation (CV). Because we leverage contextual information by labeling sequences of clauses from entire documents, for all experiments presented in this section, if not indicated otherwise, we put all instances of one document into the same fold as one sequence. Fold sizes differ slightly from each other, but folds are kept constant for all experiments.

On WikiGenerics, we carry out all three prediction tasks as defined in Section 3. On the ACE corpora, we only conduct Task NP because there are

System	generic			non-generic			macro-avg			accuracy
	P	R	F1	P	R	F1	P	R	F1	
Majority class baseline	0.0	0.0	0.0	86.8	100	92.9	43.4	50.0	46.5	86.8
Person baseline (R&F)	60.4	10.2	17.5	87.9	99.0	93.1	74.2	54.6	62.9	87.2
R&F (BayesNet)	37.7	72.0	49.5	95.0	81.9	88.0	66.4	76.9	71.3	80.6
Reimpl. (BayesNet)	38.1	67.7	48.8	94.4	83.3	88.5	66.3	75.5	70.6	81.2

Table 2: Results of **reimplemented baseline** on ACE-2 (original, unbalanced data set), 40106 instances (annotated noun phrases). Weka's stratified 10-fold cross validation, using all features.

no labels corresponding to Task Cl or Task Cl+NP.

For the experiments on WikiGenerics, we use leave-one-document-out CV, i.e., we train on 101 of the 102 documents and test on the remaining document in each fold. The total number of clauses is 10355. From ACE-2005, we use the newswire and broadcast news subsections.[4] Due to low frequency, we omit instances of NEG in our experiments, and apply a three-way classification task (GEN, SPC, USP). We present results for all remaining 40106 mentions and for the subset of 18029 subject mentions, each time using 10-fold CV.

5.2 Baseline: Local Classifier

The system for identifying generic NPs of Reiter and Frank (2010), henceforth R&F, makes use of the English ParGram LFG grammar for the XLE parser (Butt et al., 2002). As this grammar is not publicly available, we implement a similar system using exclusively the Stanford CoreNLP toolsuite (Manning et al., 2014), the Celex database of English nouns (Baayen et al., 1996) and WordNet (Fellbaum, 1999). Our system is based on dkpro (de Castilho and Gurevych, 2014). We extract the features listed in Table 1 based on the POS tags and syntactic dependencies assigned by the Stanford parser (Klein and Manning, 2002). We could not reimplement several tense- and aspect-related ParGram-specific features. In order to compensate for this, we add an additional feature (tense) with finer-grained tense and voice information, using the rules described by Loaiciga et al. (2014). Other additional features did not improve performance, which shows that R&F's set of features captures the syntactic-semantic information relevant to genericity classification quite well. Therefore, we use this feature set also for the sequence labeling model. Using the same feature set allows us to attribute any performance gain to the context-awareness of our model rather than the features.

R&F train a Bayesian network using Weka (Hall et al., 2009). The decisions of this classifier are local to each clause. They report the performance of their system on the ACE-2 corpus: Table 2 shows that the performance of our re-implemented feature set[5] is comparable to the system of R&F.[6] In all other other tables, "BayesNet R&F" refers to our re-implemented system.

R&F present the "Person baseline" as a simple informed baseline (see Table 2). We trained a J48 decision tree on this feature alone, which confirmed that only second-person mentions (the generic "you") are classified as generic, while all other mentions are classified as non-generic.

5.3 Results and Discussion

In this section, we first discuss the results of our experiments in terms of identifying generic NPs or clauses. Then we present some additional experiments testing the influence of the different feature classes and of other discourse-related information.

All tasks, WikiGenerics. The observations described in this paragraph are the same for all three prediction tasks on WikiGenerics. As Tables 3 and 4 show, our CRF models outperform the baseline system of R&F by a large margin both in terms of accuracy and F_1-score on the WikiGenerics corpus. In Task NP and Task Cl, precision and recall are quite balanced (not shown in tables). The performance of the bigram model is significantly better than that of the unigram model, increasing accuracy by about 3%, at the same time increasing F_1. In an oracle experiment, we use the previous gold label instead of the predicted one for $f_i(y_{j-1}, y_j)$, and scores increase by up to 6.6% compared to the unigram model. These results provide strong empirical evidence for our hypoth-

[4] The rest of the data comprise broadcast conversation, weblog and forum texts as well as transcribed conversational telephone, and would require specialized preprocessing.

[5] Implementation available at: `www.coli.uni-saarland.de/projects/sitent`

[6] Table 6 in Reiter and Frank's paper contains some typographical errors here. We thank Nils Reiter for making available his ARFF files, so we can provide this updated version.

System	Task NP: Genericity of Subject				Task Cl: Genericity of Clause			
	generic F1	**non-gen.** F1	**macro-avg** F1	**acc.**	**generic** F1	**non-gen.** F1	**macro-avg** F1	**acc.**
Majority class	71.9	0.0	35.9	56.1	60.3	3.7	35.1	43.7
BayesNet (R&F)	72.6	70.8	72.3	71.7	72.4	74.6	73.7	73.5
CRF (unigram)	79.3	72.6	75.9	76.4*	77.9	77.0	77.4	77.4†
CRF (bigram)	**81.3**	**76.3**	**78.8**	**79.1***	**80.8**	**80.6**	**80.7**	**80.7†**
- only clause features	79.2	71.6	75.5	76.0	79.3	78.3	78.8	78.8
- only NP features	76.8	70.8	73.8	74.1	70.7	72.6	71.8	71.7
CRF (bigram, gold)	***85.0***	***80.4***	***82.7***	***83.0***	***82.9***	***82.6***	***82.8***	***82.8***

Table 3: Results on **WikiGenerics** for **Task NP** and **Task C**. *†Difference statistically significant.

System	Task Cl+NP: Genericity of Clause (three-way)						
	GEN_gen F1	**NON-GEN_non-gen** F1	**NON-GEN_gen** F1	**macro-avg** P	R	F1	**accuracy**
Majority class	67.1	0.0	0.0	16.8	33.3	22.4	50.4
BayesNet (R&F)	69.1	69.1	26.1	54.5	58.4	56.4	65.2
CRF (unigram)	78.5	72.6	**35.4**	67.2	60.0	63.4	74.0*
CRF (bigram)	**81.3**	**76.9**	33.4	**70.3**	61.8	**65.8**	**77.4***
- two-step	80.8	75.8	28.6	61.5	**62.3**	61.9	73.4
- only clause feat.	79.4	72.6	25.3	67.0	57.2	61.8	74.3
- only NP feat.	72.9	71.4	2.5	53.0	49.9	51.4	70.0
CRF (bigram, gold)	***84.0***	***80.6***	***39.1***	***72.8***	***65.7***	***69.0***	***80.6***

Table 4: Results on **WikiGenerics** for **Task Cl+NP**. *Difference statistically significant.

esis that using context information is useful for identifying the genericity of NPs or clauses.

Task Cl+NP, WikiGenerics. In Task Cl+NP (see Table 4), only about 6% of the instances have the gold label **NON-GEN_gen** (i.e., a non-generic sentence with a generic subject), the other instances are distributed roughly evenly between the other two labels. The difficulty of Task Cl+NP thus consists in identifying this infrequent case. The three-way CRF outperforms the two-step approach both in terms of accuracy and macro-average F_1-score. The precision-recall tradeoff differs: for the **NON-GEN_gen** class, P and R of the CRF are 55.2% and 24.5% and those of the two-step-approach are 23.8% and 35.9%. The two-step approach labels more instances as **NON-GEN_gen** but does so in a less precise way. While the performance of our model leaves room for improvement on Task Cl+NP, especially with regard to the class **NON-GEN_gen**, it is worth noting that the computational model captures something about the nature of this latter class; its instances *do* look different in the feature space. The context-aware CRF using three labels performs best.

Feature set ablation. In this ablation test, shown in Tables 3 and 4, our best model (CRF bigram) uses either the set of clause-based or the set of NP-based features at a time. Clause-based features are more important than the NP-based features for all three classification tasks. An interesting observation is that the NP features alone are not able to separate the infrequent class **NON-GEN_gen** from the other two at all, the F_1-score of 2.5 shows that almost all instances of this class were labeled as one of the other two classes. In sum, this shows that whether an NP is interpreted as generic or not strongly depends on how it is used in the clause.

Task NP, ACE. Both on ACE-2 (see Table 5) and on ACE-2005 (see Table 6), the CRF outperforms the system of Reiter and Frank (2010) in terms of accuracy, and has a higher F_1-score. We give results also for subjects only as this parallels the setting of the WikiGenerics experiments (reasons for the restriction to subjects were given in Section 3). For subjects, the majority class SPC is less frequent (compare the accuracies of the two majority class baselines); only 7% of the subjects are marked as GEN, the rest are labeled as USP. The bigram model does not outperform the unigram model, but our oracle experiments show that context information is indeed useful: accuracy increases significantly and F_1 increases considerably, especially for subjects.

System	generic F1	non-generic F1	macro-avg P	macro-avg R	macro-avg F1	accuracy
Majority class	0.0	92.9	43.4	50.0	46.5	86.8
BayesNet (R&F)	47.4	87.9	65.5	**74.6**	69.8	80.4
CRF (unigrams)	49.1	93.5	75.5	68.7	71.3	88.5*
CRF (bigrams)	**51.0**	**93.7**	**76.5**	69.8	**72.4**	**88.9**
CRF (bigram, gold)	***57.6***	***94.4***	***79.8***	*73.4*	***76.0***	***90.1****

Table 5: Results on **ACE-2** for **Task NP**, 10-fold CV, folds contain complete documents. *Difference statistically significant.

System	macro-avg P	macro-avg R	macro-avg F1	accuracy
all 18029 annotated mentions				
Majority class	27.0	33.3	29.9	81.1
BayesNet (R&F)	50.8	57.2	53.8	74.5
CRF (unigram)	**61.6**	51.8	**55.1**	**83.2***
CRF (bigram)	60.6	51.7	54.8	83.0
CRF (bigram, gold)	***63.9***	***54.9***	***58.2***	***83.9****
5670 subject mentions				
Majority class	25.0	33.3	28.6	75.1
BayesNet (R&F)	51.5	**53.9**	52.7	72.5
CRF (unigram)	58.0	51.3	53.6	77.7*
CRF (bigram)	58.3	51.3	53.7	77.8
CRF (bigram, gold)	***62.4***	***56.1***	***58.6***	***79.6****

Table 6: Results on **ACE-2005** (bn+nw), **Task NP**, 10-fold CV, 3 classes: SPC, GEN, USP. *Difference statistically significant.

We identify two reasons for the fact that when evaluating on the ACE corpora, oracle information is needed to show the benefit of using bigram feature functions: (a) The frequency of **GEN** mentions in the ACE corpora is low – news contains only little generic information, so the context information is harder to leverage. (b) The ACE annotation guidelines contain some vagueness (see Section 3); this makes it harder for an automatic system to learn about regularities.

Higher-order Markov models. Another research question is whether models incorporating not only the previous label, but more preceding labels would perform even better. We turn to the Mallet toolkit (McCallum, 2002), whose CRF implementation allows for using higher-order models.[7] For example, an order-2 model considers the two previous labels. We use L1-regularization during training. Figure 1 shows that the optimum is reached for order-1 (bigram) models for each of the classification tasks for accuracy, the same tendencies were observed for F_1-score (not shown). It seems sufficient to use bigram feature functions; note that as explained in Section 4, the bigram model does not mean that only adjacent clauses influence each other – context is actually wider.

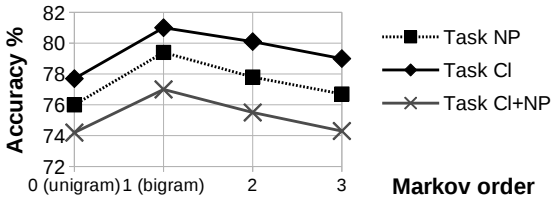

Figure 1: Labeling results for CRF models of various orders on WikiGenerics corpus.

Using coreference information. In our approximately balanced WikiGenerics corpus, 54% of all pronouns are marked as generic and 46% are marked as non-generic, which shows that there is no preference for pronouns to occur with either class. Some of the features (countability, noun type, determiner type, bare plural, and the WordNet related features) are not informative when applied to personal or relative pronouns. Sometimes, it is not even possible to determine number without referring to the antecedent (e.g., in the case of the relative pronoun 'who'). We conduct the following experiment: we automatically resolve coreference using the Stanford coreference resolution system (Raghunathan et al., 2010). We replace the NP features of each pronominal instance with the features of the first link of the coreference chain. We did not obtain a significant performance gain. One reason is that this change of features only applies to about 13% of the data. We observe that any positive changes in the classification go along with some negative changes which were often due to coreference resolution errors. One difficult step in manually annotating, and hence also in automatically resolving coreference is to determine whether a NP is generic or not (Nedoluzhko, 2013). The task of identifying generic NPs and

[7]The CRF++ toolkit, which we use in all other experiments, does not allow for higher-order models. We use CRF++ in the main experiments as it comes with a concise documentation; this helps to make our experiments easily replicable.

coreference resolution are intertwined. We plan to manually annotate at least part of our corpus with coreference information in order to test to what extent the classification of the pronouns' genericity status can profit from including antecedent information.

6 Conclusion

We have presented a novel method for labeling sequences of clauses or their subjects with regard to their genericity, showing that genericity should be treated as a discourse-sensitive phenomenon. Our experiments prove that context information improves automatic labeling results, and that our model outperforms previous approaches by a large margin.

The major contributions of this work include the study of genericity both on the NP- and clause-level, and the study of the interaction of these two levels. Our results of Task Cl+NP show that our model indeed captures the three different types of clauses resulting from the combination of NP-level and clause-level genericity.

During the development of our annotation scheme, we found that it is beneficial to focus on genericity, disentangling it from the issue of specificity. Our work provides a step forward to finding reliable ways to apply semantic theories of genericity in practice, and we also provide a new state-of-the-art system for automatically labeling generic expressions. This in turn lays foundations for natural language processing tasks requiring text understanding.

Future Work. Our present approach for annotating and automatically classifying targets the subjects of each clause. We have not attempted to tackle the task of classifying the genericity status of other dependents, as they are even harder to classify than subjects, and a concise annotation scheme has to be worked out in order achieve an acceptable inter-annotator agreement on this task. Another related distinction is the one between habitual, stative and episodic sentences (Mathew and Katz, 2009), which applies to both what we call generic and non-generic sentences. No large corpora exist to date, but studying the interaction of these phenomena is on our research agenda.

Acknowledgments

We thank the anonymous reviewers, Alexis Palmer, Nils Reiter and Melissa Peate Sørensen for their helpful comments related to this work, and our annotators Christine Bocionek and Kleo-Isidora Mavridou. This research was supported in part by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative (DFG), and the first author is supported by an IBM PhD Fellowship.

References

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. CELEX2. Philadelphia: Linguistic Data Consortium.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *Proceedings of the 2002 workshop on Grammar engineering and evaluation-Volume 15*, pages 1–7. Association for Computational Linguistics.

Gregory Norman Carlson. 1977. *Reference to kinds in English*. Ph.D. thesis.

Gregory N. Carlson. 2005. Generics, Habituals and Iteratives. In Alex Barber, editor, *Encyclopedia of Language and Linguistics*. Elsevier.

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 (OIAF4HLT) at COLING*, pages 1–11.

Christiane Fellbaum. 1999. *WordNet*. Wiley Online Library.

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

Annemarie Friedrich and Alexis Palmer. 2014a. Centering Theory in natural text: a large-scale corpus study. In *Proceedings of KONVENS 2014*. Universitätsbibliothek Hildesheim.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. *In Proceedings of the Linguistic Annotation Workshop VIII*, page 149.

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.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Aurelie Herbelot and Ann Copestake. 2009. Annotating genericity: How do humans decide? (A case study in ontology extraction). *Studies in Generative Grammar 101*, page 103.

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*, pages 3–10.

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.

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 '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Godehard Link. 1995. Generic information and dependent generics. *The Generic Book*, pages 358–382.

Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French Verb Phrase Alignment in Europarl. In *Proceedings of LREC 2014*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Thomas A. Mathew and E. Graham Katz. 2009. Supervised Categorization of Habitual and Episodic Sentences. In *Sixth Midwest Computational Linguistics Colloquium*, Bloomington, Indiana: Indiana University.

Andrew K McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. 2003. ACE-2 Version 1.0 LDC2003T11. Philadelphia: Linguistic Data Consortium.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111.

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*, page 896.

Massimo Poesio and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In *LREC*.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Nils Reiter and Anette Frank. 2010. Identifying Generic Noun Phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49, Uppsala, Sweden, July. Association for Computational Linguistics.

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-Volume 1*, pages 149–156. Association for Computational Linguistics.

Sangweon Suh, Harry Halpin, and Ewan Klein. 2006. Extracting common sense knowledge from wikipedia. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at ISWC*, volume 6.

Sangweon Suh. 2006. Extracting Generic Statements for the Semantic Web. *Master's thesis, University of Edinburgh*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus LDC2006T06. Philadelphia: Linguistic Data Consortium.