Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, January 2012

Semantic Annotation for the Digital Humanities

– Using Markov Logic Networks for Annotation Consistency Control

Anette Frank, Thomas Bögel, Oliver Hellwig*, Nils Reiter

Published by CSLI Publications

LiLT volume 7, issue 8 January 2012

Semantic Annotation for the Digital Humanities

– Using Markov Logic Networks for Annotation Consistency Control

Anette Frank, Thomas Bögel, Oliver Hellwig*,
Nils Reiter, *Department of Computational Linguistics, Heidelberg University, *South Asia Institute, Heidelberg University*

Abstract

This contribution investigates novel techniques for error detection in automatic semantic annotations, as an attempt to reconcile error-prone NLP processing with high quality standards required for empirical research in Digital Humanities. We demonstrate the state-of-the-art performance of semantic NLP systems on a corpus of ritual texts and report performance gains we obtain using domain adaptation techniques. Our main contribution is to explore new techniques for annotation consistency control, as an attempt to reconcile error-prone NLP processing with high quality requirements. The novelty of our approach lies in its attempt to leverage multi-level semantic annotations by defining interaction constraints between local word-level semantic annotations and global discourse-level annotations. These constraints are defined using Markov Logic Networks, a logical formalism for statistical relational inference that allows for violable constraints. We report first results.

LiLT Volume 7, Issue 8, January 2012.
Semantic Annotation for the Digital Humanities.

1 Introduction

The work described in this paper is embedded in an interdisciplinary project that aims at analyzing regularities and variances in the event structures of Nepalese rituals.[1] The focus of this project is on investigating the event structure of rituals by applying computational linguistic analysis techniques to written descriptions of rituals.

For scholars working in applied research in Digital Humanities, it is important that any evidence derived from computational analysis is accurate and reliable. Thus, for our project – as for others with similar foundations – it is of utmost importance to produce high-quality automatic annotations. But despite the many prospects that computational analysis can offer for empirical research in the Humanities, assuring near-to-perfect quality of computational analysis is still beyond the limits of state of the art systems.

As has been shown in current work on domain adaptation (e.g. Daumé III (2007)) including our own, there is potential in improving the quality of current NLP tools by applying domain adaptation techniques. However, the gap in performance between current system outputs and (near-to-)perfect annotation quality is still considerable.

In this contribution we investigate novel techniques for annotation error detection to guide manual annotation control or to acquire training material for domain adaptation. In contrast to most earlier work that concentrates on detection of part of speech (PoS) or parsing errors, our focus is on semantic annotation. The novelty of our approach lies in its attempt to leverage multi-level semantic annotation for annotation consistency control and error detection. We will interface *local* word-level semantic annotations with *global* discourse-level annotations. Concretely, we will define interaction constraints between annotations produced by a word sense disambiguation (WSD) and a coreference resolution (CR) system. These constraints will be defined using Markov Logic Networks (MLN, Richardson and Domingos (2006)), a first-order predicate logic formalism for statistical relational inference that allows the definition of violable constraints.

The paper is organized as follows. Section 2 reviews previous work on error detection for linguistic annotations. Section 3 presents an evaluation of the performance of various semantic processors: word sense disambiguation (WSD), frame-semantic labeling (SRL) and coreference resolution (CR) systems, which we adapted to the domain of ritual

[1] The project is part of the collaborative research center (Sonderforschungsbereich, SFB) "SFB 619: Ritual Dynamics" at Heidelberg University; http://www.ritualdynamik.de.

texts. Section 4 motivates our approach for cross-level semantic annotation consistency control and introduces a method for consistency checking using Markov Logic Networks (MLNs). Section 5 reports the results of our first experiments for annotation error detection. Section 6 summarizes our findings.

2 Related Work

Methods for detecting annotation errors have been developed early on in the context of treebank construction, to enhance the quality of linguistic annotations for training supervised systems for PoS tagging or parsing (see e.g. Dickinson and Meurers (2003)). Prevalent techniques include observations based on corpus statistics, such as checking for deviations in PoS assignments over identical n-grams in a given corpus (Dickinson and Meurers, 2003, Loftsson, 2009), or inferring infrequent or "negative" n-grams from clean corpora to detect PoS annotation errors (Květoň and Oliva, 2002). Other techniques make use of manually defined or learned context-sensitive error detection (and correction) rules (Dickinson and Meurers, 2003).

Detecting errors in syntactically annotated corpora works in similar ways, by extracting grammar rules or trees from labeled corpora and comparing the obtained rules or structures and their frequencies to those obtained from validated annotated corpora. Methods range from comparing full or partial structures to trees including surface frontiers, to the use of strict or partial overlap criteria (cf. e.g. Dickinson (2010)).

There is little research, to date, that investigates methods for detecting errors in semantic annotation. Given the difficulty of the task, most annotation projects rely on the four-eye principle to detect disagreements among annotators. Yu et al. (2008) are the first to investigate methods for detecting *mistaken agreements* between annotators in assigning word senses. Here, all agreed-upon annotations are compared against the assignments of a supervised WSD system. Clearly, this method requires a substantial amount of annotated instances for training the WSD system, thus it is only suited for high-volume annotations for known lexical items.Yu et al. (2008)'s system identifies 40% of the erroneous annotations in the data. This performance is insufficient for fully automatic error correction. But it achieves a cost-effectiveness ratio that seems high enough to propose suspicious instances for manual control.

Dickinson and Lee (2008) apply a data-driven approach to detect errors in predicate-argument structure annotations that relates to earlier work using n-gram error detection for identifying syntax errors.

Basically, the method identifies identical pieces of text with variational annotation. It thus requires a substantial amount of labeled data with overlapping surface strings.

To our knowledge, there is no prior work that makes use of multiple annotation layers to detect inconsistencies in manual or automatic semantic annotations. Also, no attempts have been made to integrate statistical observations with logical constraints to define inter-level dependencies between annotations for this purpose. In our work, this will be attempted using Markov Logic networks, as explained in Section 4.

3 Multi-Level Semantic Annotation

This section describes the textual data, the preprocessing steps and the individual semantic analysis components we use for analyzing the event structure of rituals, including the performance they achieve when applied to ritual texts.

3.1 Corpus of ritual descriptions

Our corpus of ritual texts consists of ritual descriptions obtained from two types of sources: part of the corpus is supplied from modern ethnographic observations of rituals in Nepal, another from Sanskrit manuals about Nepalese rituals, which are translated to English by ritual experts.[2] The complete ritual corpus currently consists of 48 descriptions of rituals, ranging from 3 to 339 sentences and comprises 97,460 tokens. While most texts deal with *rites de passage*[3] in Nepal and are, therefore, rather consistent at the topic level, there are clear differences in their language styles. The translations from Sanskrit texts consist mainly of short sentences with an average sentence length of 18 words. They frequently use a terse, condensed language with many imperatives and nominal constructions, which reflects the style of the underlying Sanskrit originals:

"Now, the rules for the ricefeeding ceremony."
"Hand over the flower basket."

The style of the ethnographic descriptions may be characterized as "scientific prose" with longer sentences and nested substructures (average

[2]Most texts are drawn from the works of Gutschow and Michaels (2008) and Gutschow and Michaels (2005). An extension of the current corpus is planned on the basis of the upcoming volume about marriage rituals.

[3]The term *rite de passage* denotes rituals that are performed during the transition between important states in a person's life. Our ritual corpus comprises descriptions of classical Indian transitory rituals (*saṃskāra*, e.g., first feeding of solid food and beginning of the Veda study) as well as typical Nepalese rituals such as the "Marriage to the Ihi fruit".

sentence length: 26 words):

> "The pair of pots designated for a female spirit is likewise painted and the lump of clay worshipped as Śiva or Agni during the Buddhists' Girl's Marriage to the bel fruit fashioned out of clay."

Both text types contain numerous terms that are specific for South-Asian material and religious culture such as names of gods (*Śiva, Agni*) or indigenous fruits (*bel fruit*). In order to facilitate processing, these terms are replaced by approximate translations to English provided by ritual experts prior to processing. Since we store the original terms as annotations, we can re-insert them after processing (cf. Reiter et al. (2011) for further details on the text characteristics of this corpus).

3.2 NLP architecture

All systems are integrated in a full-fledged natural language processing architecture based on UIMA[4], with analysis results stored as stand-off annotations. The architecture comprises various processors for the major preprocessing steps: tokenization, PoS tagging, lemmatization and chunking, as well as the semantic and discourse-level analysis components discussed below: word sense disambiguation, semantic role labeling and coreference resolution.

Given the low quality we obtained using off-the-shelf processors trained on common text genres (i.e., newspaper), we experimented with various domain adaptation techniques for PoS tagging and chunking (cf. Reiter et al. (2011) for details).

For **PoS tagging**, a standard model trained on the Wall Street Journal (WSJ) achieved an accuracy of approximately 90% on a manually annotated test set consisting of 672 sentences. We were able to improve on that by retraining the model on the concatenation of the WSJ and oversampled ritual data. This way, the PoS tagger achieved a performance of just under 97%. In a similar fashion, we improved the model for **chunking** from an f-score of 0.86 (when trained on the WSJ) to an f-score of over 0.88 when trained on the concatenation of the WSJ and oversampled ritual data.

In the following, we discuss processing components for three levels of semantic and discourse-level annotation and their adaptation to the ritual domain: word sense disambiguation using WordNet (Fellbaum, 1998) as sense inventory; semantic role labeling based on FrameNet (Baker et al., 1998) and coreference resolution.

[4] `http://uima.apache.org`

		MFS	$\text{UKB}_{\text{WN 2.0}}$	$\text{UKB}_{+\text{rit-node}}$
Nouns	Coverage	94.5	93.3	93.3
	Precision	59.8	60.2	**64.1**
	Recall	**60.0**	53.7	57.3
	F-Score	59.9	56.8	**60.5**
Adjectives	Coverage	88.4	86.9	86.9
	Precision	48.3	**51.2**	49.8
	Recall	**49.3**	**49.3**	47.8
	F-Score	48.8	**50.2**	48.8
All Words	Coverage	94.3	93.1	93.1
	Precision	53.9	54.2	**56.4**
	Recall	**54.5**	49.9	51.8
	F-Score	**54.2**	51.9	54.0

TABLE 1 Evaluation results: WSD without and with domain adaptation

3.3 Word Sense Disambiguation

For word sense annotation we employ the graph-based UKB system of Agirre and Soroa (2009). While supervised WSD systems rely on manually labelled training data, UKB explores the graph structure of WordNet using the PageRank algorithm for assigning senses to target words in a given context.

WSD performance. To build a gold standard for testing UKB's performance, we randomly chose 50 sentences from all ritual descriptions. These sentences were annotated independently by two annotators with word senses from WordNet 2.0. Both annotators have a computational linguistics background. Differences between the two annotations have been adjudicated.[5] This resulted in 462 annotated nouns, verbs, adjectives and adverbs, forming our gold standard for WSD.

We assessed the performance of UKB using precision and recall as evaluation metrics, calculated for individual word types and micro-averaged over all types. As the semantic annotation of verbs will be mainly covered by FrameNet annotations, we specifically report on the performance of WordNet sense disambiguation for nouns and adjectives, next to performance on all words. Here and in all the following experiments, the WSD system selects candidate synsets based on the PoS tags provided by our own domain-adapted, probabilistic PoS tagger.

[5] In two cases WordNet 2.0 did not contain appropriate concepts for annotation: "*bel* fruit" (Sanskrit *bilva*; a fruit used for worshipping Śiva) and "*block* print". These words were left unannotated.

The performance results for different system configurations are summarized in Table 1. We assigned the most frequent sense (MFS) from WordNet 2.0 as a baseline. This baseline achieves a precision of 53.9% and a recall of 54.5% for all words. For 5.7% of the tokens, the baseline implementation does not return a word sense. This loss in coverage is mainly caused by erroneous PoS assignments.

We first tested the performance of UKB 0.1.6 using standard WordNet (2.0). The system achieves a precision of 54.2% and a recall of 49.9% (for all words) and thus performs below the MFS baseline (the loss in recall outranks the gain in precision), which is not unusual for unsupervised WSD systems. The coverage drops by a small amount to 93.1%.

Domain adaptation for WSD. In order to adapt UKB to the ritual domain, we enriched the WordNet database with domain-specific sense information. We acquired senses that may be characteristic for the ritual domain from a Digital Corpus of Sanskrit (DCS, Hellwig (2011)). This corpus is designed as a general-purpose philological resource that covers Sanskrit texts from 500 BCE until 1900 CE without any special focus on the ritual domain. In this corpus, approximately 400,000 tokens have been manually annotated with word senses from WordNet 2.0. Using this annotated corpus for domain sense acquisition was motivated by the supposition that even general passages from Sanskrit literature may contain a significant amount of senses that are relevant for the ritual domain.

We linked all 3,294 word senses that were annotated in this corpus to a newly introduced non-lexicalized pseudo-synset `rit-topic`. As UKB calculates the page rank between sense-related words in the WordNet database, introducing this node increases the chances that senses specific for Nepalese culture receive a higher rank (cf. Reddy et al. (2010) for a similar approach).

As seen in Table 1, linking domain-related senses to a pseudo-synset results in an improvement of 2.2 points in precision and 1.9 points in recall for all words, when compared to $\text{UKB}_{\text{WN2.0}}$. Moreover, the domain-adapted UKB system now closely matches the MFS baseline in F-Score. Note further that for nouns the domain-adapted WSD system obtains the best results (P: 64.1%, F: 60.5), and outperforms the MFS baseline in terms of precision (+4.3) and f-score (+0.6), with only a slight loss in recall (R: 57.3%; -2.7) and coverage remaining stable. This is in line with our general aim towards producing precise annotations.

3.4 Frame-semantic labeling

Semi-automatic frame-semantic labeling. We added frame semantic annotation to the ritual descriptions in a semi-automatic way. First, a learner trained on small amounts of annotated data was used to assign frames in unannotated descriptions. The assigned frames were checked by two annotators, and differences were adjudicated by one supervisor. In a second step, semantic roles were assigned manually to the adjudicated frames by two annotators, and were again checked for consistency by the supervisor.

Depending on the complexity and the ambiguity of a frame, we observed an inter-annotator agreement between $\kappa = 0.619$ (frame Manipulation) and $\kappa = 1.0$ (frame Cutting) for frame annotation. For role annotation, we observed a global $\kappa = 0.469$, which indicates rather low agreement. However, a closer look at the data reveals that 89.4% of the differences in role annotations occur when one annotator annotates a role that the other annotator does not recognize.

Using this double annotation approach, we built up a manually checked gold corpus that contains 1,505 frames of 12 different types and 3,061 roles of 94 different types.

Automatic semantic annotation quality. To reduce the need for time-consuming manual annotation, we experimented with existing semantic role labeling systems. We evaluated the probabilistic SRL system Semafor (Das et al., 2010), which has been trained with FrameNet (1.5) data, against the manually annotated gold corpus described above.

Semafor achieved P: 49.6%, R: 34.4% and F: 40.6 for frame labeling.[6] Error analysis shows that the accuracy of Semafor varies strongly depending on the frames. Semafor performs poorly with frames that carry culture-specific notions or are evoked by unusual lexemes in the ritual descriptions. For the frame text_creation, for instance, Semafor yields R: 0.2%, P: 0.9% and F: 0.3, because it labels target words such as *chant* consistently with the frame Communication_manner, while our group decided to annotate the frame Text_creation in these cases.[7] The low recall can be explained by the fact that verbs such as *recite*, which are missing in FrameNet, are annotated manually with the frame Text_creation in the gold corpus. On the other hand, we observe good accuracy for less specialized frames such as Placing (P: 82.2%, R: 76.2%, F: 79.1). An analysis of coverage gaps

[6] Eight cases with multiword gold targets were excluded from consideration in automatic evaluation, as it is unclear whether partial matches can be considered as meaning preserving.

[7] Chantings in rituals are usually not meant as a form of communication.

according to Palmer and Sporleder (2010) shows that about 75% of all errors in frame assignment are caused by insufficient training material in FrameNet.[8]

The evaluation of semantic roles was restricted to the roles of those frames that were annotated correctly by Semafor. On these 1182 roles, Semafor achieved P: 58.2%, R: 62.1% and F: 60.1, allowing both partial and perfect overlap of spans; P: 52.0%, R: 55.5%, F: 53.7 if restricted to perfect match.[9] As major sources of error, we identified non-local roles and non-core roles that are missing in Semafor's output, domain specific vocabulary of our texts, and syntactic peculiarities such as imperatives. On the whole, we are confident that system annotations for frames and roles can be improved by retraining Semafor on our labeled domain data.

3.5 Coreference Resolution

Coreference Resolution using BART. We chose BART (Versley et al., 2008) as our primary tool for coreference resolution. BART implements a classical approach towards coreference resolution based on a classification of mention-pairs, as described in Soon et al. (2001). Integrated preprocessing components (PoS tagging, constituent parsing, etc.) are used to extract mentions and their features. The system includes precompiled models for anaphora and coreference resolution using a standard feature set for pair-wise classification trained on the MUC6 data set (Chinchor and Sundheim, 2003). Best results were achieved using the precompiled MaxEnt model.

Domain adaptation techniques. Given extremely poor results when using BART as off-the-shelf coreference resolver (cf. Reiter et al. (2011)), we tested several strategies to enhance its performance on ritual texts.

First, to reduce noise from preprocessing, we adapted BART's integrated preprocessing pipeline to include our own domain-adapted components for PoS tagging and chunking.

Two further enhancements are used to tailor the system to our targeted domain and interests. (i) After mention detection, a WordNet lookup filters out mentions of specific semantic classes. This allows us

[8] Using the notation introduced in (Palmer and Sporleder, 2010, p. 932f), the detailed numbers are as follows: NOTR-LU: 8.5% (83 instances; including those cases in which the annotation report of FrameNet gives less than three annotated instances), NOTR-TGT: 10.1% (99), UNDEF-LU: 17.7% (174), UNDEF-TGT: 36.9% (362).

[9] Precision could be slightly underestimated due to a number of roles (80) in Semafor's output that are not annotated in the gold standard, but could still be correct.

	MUC			B^3		
	P	R	F	P	R	F
BART preprocessing	37.68	59.77	46.22	28.79	46.28	35.5
UIMA preprocessing	41.9	60.53	**49.52**	32.21	46.81	**38.16**

TABLE 2 Standard vs. domain-adapted preprocessing pipeline

to concentrate on the most important and most frequent entity types: persons and gods (as opposed to non-animated objects). Also, (ii) we included domain-specific knowledge to improve the predictions of BART's semantic agreement features: We extended BART's internal database for names and procedures with a new category for gods and added gender information for items frequently occurring in ritual texts to the existing knowledge databases.

Evaluating automatic coreference annotation quality. We evaluated BART's performance on manually annotated gold standards using MUC and B^3 as evaluation metrics. Both metrics compare chains of mentions produced by the system with corresponding chains in the gold standard to measure coreference resolution performance. MUC counts *missing links* between mentions in the system's output relative to the gold standard (cf. Vilain et al. (1995)). Despite known shortcomings of MUC, it is still widely used. Bagga and Baldwin (1998) resolve these issues with the introduction of the B^3 metric that judges *each mention individually*, resulting in a stricter and more realistic evaluation metric for most scenarios.

Evaluation results: processing pipelines. We tested different pipeline architectures, using our own domain-adapted tagger and chunker (UIMA pipeline) in contrast to BART's pipeline that includes a standard model for full parsing using the Stanford parser (cf. Table 2).

Using chunks provided by the UIMA pipeline clearly outperforms BART's internal pipeline across both evaluation metrics. Given these results, we chose to use the UIMA pipeline for preprocessing for all subsequent experiments.

Evaluation results: entity subtypes and domain knowledge. We evaluated the two domain-specific adaptions discussed above: (i) restricting coreference resolution to entity subtypes, and (ii) extending BART's semantic knowledge by adding gender information and semantic categories for frequently occurring terms.

(i) Table 3 shows overall performance improvements for *restriction* to the entity types *person* and *god*. Gains are very high for MUC, while moderate and mostly oriented towards precision for B^3. This holds both

	entity subtypes	MUC P	MUC R	MUC F	B^3 P	B^3 R	B^3 F
Standard model	all	41.9	60.53	49.52	32.21	**46.81**	38.16
	person	**55.73**	**64.15**	**59.64**	**36.82**	40.46	**38.56**
	object	25.78	58.41	35.86	23.74	58.2	33.73
Domain gender model	all	39.9	**62.82**	48.8	29.26	**50.51**	37.05
	person	55.59	63.89	59.62	**36.92**	40.2	38.49
	object	25.6	**61.83**	**36.13**	**23.80**	**61.5**	**34.32**

TABLE 3 Results for CR with entity type restrictions and gender database

	P	R	F
Standard model	**49.1**	78.9	**60.5**
Domain gender model	47.7	**80.4**	59.9

TABLE 4 Identification of mentions

for the standard gender model of BART (upper part) and the domain-adapted model (lower part). This result fits well with our main interest in analyzing event chains from rituals, where coreference information for the main actors is of primary importance, and our general interest in achieving high-quality annotations.

(ii) For the domain-specific enhancements to the *gender model*, we observe clear gains in recall in comparison to the non-adapted model.[10] This goes along with a slight drop in precision across all categories. For the *person* entity subtype, however, the gender model does not have a clear impact, with pretty stable recall and precision. Table 4 shows that in the model with enhanced gender features more mentions can be linked to entity chains (recall of mention identification rises to 80.4%). This explains the general improvement in recall, with a trend towards a drop in precision, due to misclassifications. In this respect, the person entity type shows robust behavior, with almost identical overall performance. We may still expect improved performance of the domain model when analyzing larger data sets.

Overall, our *person-restricted domain-adapted models* achieve clearly improved precision, with a boost of 18.05 points (`MUC`) and 8.16 points (B^3) when compared to the unadapted standard BART model (cf. Table 2), with solid gains in f-scores (8.13 and 3.06 points, respectively).

[10]Table 3 (domain gender model) highlights results that outperform the corresponding variant of the standard model.

4 Exploiting Multiple Layers for Consistency Control

As seen above, we can achieve significant improvements in labeling accuracy for WSD and CR by applying different domain adaptation strategies. For frame-semantic annotation, we identified issues of domain-specific senses that can be addressed by retraining Semafor on the domain corpora that were labeled semi-automatically.

Still, it soon became clear in our interdisciplinary project that for the ritual scientists it is crucial that any observations obtained from data analysis are reliable. As we have seen, this cannot be realistically achieved by the current state of the art in NLP. Manual annotation, on the other hand, seems out of reach for a substantial amount of data.

As a way to counterbalance error-prone automatic annotation with measures to ensure high annotation quality, we investigated methods for consistency control that can help identify erroneous annotations in the data, for *targeted manual correction* or to *acquire valuable training data for improving automatic labelers*. As outlined in Section 2, methods for error detection have by now concentrated on morphological and syntactic analysis. The few attempts reported on consistency checking in semantics are confined to a single level of annotation (Yu et al., 2008) or mainly draw on techniques for syntactic error detection (Dickinson and Lee, 2008). Our work focuses on error detection techniques that leverage multiple levels of (discourse-)semantic annotation.

Intra- and inter-level consistency. In general, consistency control can be addressed from two perspectives: relying on evidence obtained for a single level of annotation, or else by deriving consistency constraints from known interactions or dependencies across levels that can be used to detect outliers in annotations. We refer to these opposing views as intra- and inter-level consistency.

Classical methods for **intra-level** consistency control are *voting* or *classifier combination* using alternative labelers. This is a well-known, effective technique for improved system results in generic classification tasks. It is evaluated in Loftsson (2009) for PoS error detection and could be applied to any level of analysis, including semantics. Other methods rely on frequency distributions obtained from corpora.

The focus of our work is on **inter-level** consistency control. In particular, we exploit dependencies between *local* and *global* annotation decisions, by interfacing word-level and discourse-level semantic annotation.

Discourse-level semantic dependencies. Our approach starts from a discourse perspective and the observation that coherence at the discourse level affects disambiguation decisions that are typically

taken at the word or sentence level, such as WSD or SRL. This dependency is at the heart of the *one-sense-per-discourse (OSD)* hypothesis (Gale et al., 1992) that was successfully exploited for WSD (Yarowsky, 1995).

As we focus on the semantic annotation of discourse in the form of ritual descriptions, we can exploit discourse-level constraints for semantic annotation and vice versa, to detect erroneous annotations. Specifically, we will exploit dependencies between coreference resolution (CR) and word sense disambiguation (WSD).

CR establishes coreference chains, consisting of a set of so-called mentions, typically common nouns, pronouns or proper names. This set is also referred to as a *(discourse) entity*, as all mentions jointly refer to a single entity. The task of **WSD** is to select a specific sense from the set of possible senses of a word that is appropriate in the given context. A natural assumption for the *dependency between CR and WSD* is that all common nouns contained in an entity are closely sense-related. Following the OSD hypothesis, this should be trivially true for multiple occurrences of the same common noun. For lexically distinct nouns, we can still assume that for coreferring, but ambiguous nouns, their contextually correct senses are closely related.

We will test this hypothesis by defining two **consistency constraints** that determine *sense selection* and the *assignment of mentions to a discourse entity*. They predict:

Cons.ws: for a mention m in a given entity e, sense selection (i.e. WSD) chooses a sense s that is close to a "central" concept representation c for entity e.

Cons.cr: for a given mention m with contextually assigned sense s, m is assigned to an entity e whose "central" concept c is closely related to or compatible with s.

We compute such a central or "centroid" semantic representation for discourse entities using the graph-theoretical notion of a *key player*. The key player is a measure that determines a central node in a graph by choosing a single node that is closest to all other nodes (Navigli and Lapata, 2010). In our case, we compute a *key player sense* for an entity from a semantic graph we build from all word senses of all its mentions, using WordNet. The edges of the graph correspond to the sense relations defined in WordNet, choosing the shortest distances between connected senses (cf. Bögel (2011)).

We then use the distance[11] d between word senses s and the key

[11]The distance measure is based on counting edges between nodes. Of course, other measures of similarity or distance could be used.

```
// Declarations, used in all rule sets
has_sense(ment,sen!) // each mention is assigned exactly one sense
poss_sense(ment,sen) // mentions have possible senses
in_m_e(ment,ent!)    // each mention is assigned to exactly one entity
centroid(ent,cen!)   // each entity has exactly one centroid
dist(sen,cen,int)    // distance betw. sense and centroid in path length
```

FIGURE 1 Predicate declarations common to all rule sets

player sense c of an entity to estimate the consistency of sense assignment to a mention and assignment of a mention to an entity, according to the constraints defined above:

If d is small, **Cons.ws** predicts sense assignment s to a mention m in e to be consistent with discourse-level decisions captured in an existing entity e. If an alternative sense s' is closer to c, the decision of the WSD system needs to be revised, and s' is a candidate sense to consider for m. If, on the other hand, **Cons.cr** finds a mention m in e whose assigned sense s is not close enough to c or closer to the centroid c' of another entity e', the decision of the CR system needs to be revised, and e' is a candidate entity to consider for m.

That is, we can compute semantic distances between assigned or possible senses of a mention and the centroid concepts of established discourse entities to detect violations of the coherence constraints. Any instances that incur (more or less severe) violations of these constraints should point us to outliers in semantic annotations.

Defining dependencies using Markov Logic Networks. Markov Logic is a formalism that uses weighted first-order predicate logic formulas. Formulas with a low weight are "cheaper" to violate, while the violation of formulas with higher weights is more expensive. Weights can be specified manually or learned from data. Next to weighted formulas, the system also allows the encoding of "hard rules" (or hard constraints) that cannot be violated. They receive a very high weight (∞).

Figures 1 to 4 show the declarations and rule sets we defined to implement the constraints formulated above. All rules and definitions follow the syntax of Alchemy[12], with $!p$ denoting negation of a literal p, $c!$ enforcing uniqueness on a variable assignment for c, and $+c$ enforcing estimation of weights over formulas of individually grounded variables c.

We define three variants of rule sets (Sets I, II and III) for two types of inference rules: one targeted at predicting assignment of a mention to an entity according to **Cons.cr**, the other targeted at predicting assignment of a sense to a mention m, according to **Cons.ws**. That is,

[12] `http://alchemy.cs.washington.edu`

```
// CR
has_sense(m,s) ^ centroid(e,c) ^ dist(s,c,d) ^ d <= dx => in_m_e(m,e)
has_sense(m,s) ^ centroid(e,c) ^ dist(s,c,d) ^ d > dx => !in_m_e(m,e)

// WSD
poss_sense(m,s) ^ in_m_e(m,e) ^ centroid(e,c) ^ dist(s,c,d) ^ d <= dx
                                                    => has_sense(m,s)
poss_sense(m,s) ^ in_m_e(m,e) ^ centroid(e,c) ^ dist(s,c,d) ^ d > dx
                                                   => !has_sense(m,s)
```

FIGURE 2 Rule Set I for CR and WSD, $dx = 0, \ldots, n$

```
// CR
has_sense(m,s) ^ centroid(e,c) ^ distance(s,c,+d)  => in_m_e(m,e)
has_sense(m,s) ^ centroid(e,c) ^ distance(s,c,+d)  => !in_m_e(m,e)

// WSD
poss_sense(m,s) ^ in_m_e(m,e) ^ centroid(e,c) ^ distance(s,c,+d)
                                                    => has_sense(m,s)
```

FIGURE 3 Rule Set II for CR and WSD

we define the target predicate `in_m_e` in the first case, and the predicate `has_sense` in the second.

Figure 1 displays the modeling predicates in the declaration part used for all rule sets.

Set I (Figure 2) makes use of a distance threshold ($dx = 0, \ldots, 3$) between m's (possible) senses and the entity centroid c for the assignment of a possible sense s to a mention (Cons.ws) or a mention m to an entity (Cons.cr).

In order to determine plausible distance thresholds, ***Set II*** (Figure 3) learns rule weights for the entire range of individual distances observed in our data ($0-30$), for both positive and negative assignments. Rules for Set II are similar to Set I, but specify individual distances `distance(s,c,+d)`. Due to the plus sign, the rules are compiled to individual rules for each distance value and thus, for each distance we obtain an individual rule weight.

Set III (Figure 4) offers a formulation that does not resort to a fixed distance threshold or a spread of distinct distance ranges as in *Set I* and *Set II*. Instead, we define a discriminative rule set that assigns a mention to an entity e' if the distance of its sense s to the centroid of e' is smaller than the distance between s and the centroid of the automatically established entity e; otherwise, assignment of m to e is

```
// Predicates only used in set III
inferred_in_m_e(ment, ent!)
inferred_has_sense(ment, sen!)

// CR
// 1. if a mention is closer to another entity's centroid, change decision
in_m_e(m,e) ^ has_sense(m,s) ^ centroid(e,c) ^ distance(s,c,d) ^
centroid(e',c') ^ e != e' ^ distance(s,c',d') ^ d' < d
                                                  => inferred_in_m_e(m,e')

// 2. if mention isn't closer to another entity's centroid, keep decision
in_m_e(m,e) ^ has_sense(m,s) ^ centroid(e,c) ^ distance(s,c,d) ^
centroid(e',c') ^ e != e' ^ distance(s,c',d') ^ d' >= d
                                                   => inferred_in_m_e(m,e)

// WSD
// 1. if a sense is closer to centroid than another sense, keep decision
in_m_e(m,e) ^ has_sense(m,s) ^ poss_sense(m,s') ^ s != s' ^
centroid(e,c) ^ distance(s,c,d) ^ distance(s',c,d') ^ d <= d'
                                                 => inferred_has_sense(m,s)

// 2. if another sense is closer to the centroid, change the decision
in_m_e(m,e) ^ has_sense(m,s) ^ poss_sense(m,s') ^ s != s' ^
centroid(e,c) ^ distance(s,c,d) ^ distance(s',c,d') ^ d > d'
                                               => inferred_has_sense(m,s')
```

FIGURE 4 Rule Set III for CR and WSD

preserved. In a similar way we define (alternative) assignment of a sense s' to a mention m if the distance between the mention's entity centroid c and s' is smaller than the distance to an alternative sense s.

All the above rules are defined as soft constraints. A small number of hard constraints model the WSD and CR task, stating, e.g., that each mention is assigned a single sense and assigned to a single entity. In addition, we experimented with a variant of Set III, ***Set III***$_{hard}$, in which the same rules have been defined as hard constraints.

5 Experiments and Evaluation

Data processing. We processed ritual descriptions using UKB$_{+\text{rit-node}}$ as a WSD system and BART in its best configuration (gender model restricted to persons/gods) as a coreference resolution system. For this task, entities were filtered to only contain common nouns. This way, centroids are sharply defined, being entirely based on nominal senses. We exported the resulting data into a collection of MLN predicates. A small set of data has been annotated manually to serve as development and test sets. Table 5 shows an overview of the data sets we used.

WSD	train	dev	test
# tokens	2656	53	141
# types	156	11	22
tokens/type	19.05	4.8	6.4
sense ambiguity (chain) (avg/median)	2.92/2	2.21/2	2.56/2
sense ambiguity (all nouns) (avg/median)	3.62/3	3.38/3	3.5/3

CR	train	dev	test
# mentions	3795	41	78
# chains (entities)	602	14	19
# NNs/chain	6.18	2.86	4.05

TABLE 5 Information about training, development and test data sets

Evaluation measures. Since our main goal is error detection, we report precision, recall and f-score for the detection of mistakes in automatic annotations. Ideally, we want high precision (i.e. small number of false positives) and high recall (i.e. small number of false negatives), to be able to propose potential annotation mistakes for manual control. To gain better insight into the data, we further evaluated classical performance measures for WSD and CR for the inferred against gold annotations using the MLN constraints for the best rule set.

Experiments. We evaluated the performance of the consistency constraints defined above in a number of experiments, using the Alchemy implementation of Markov Logic. We first determined plausible distance thresholds for Set I by inducing rule weights for individual distances (Set II) on a large set of automatic annotations. Based on this, we selected $dx = 0, \ldots, 3$ for WSD and $dx = 0, \ldots, 2$ for CR in Set I. Running evaluation experiments on the development data sets for Sets I, II and III did not yield significant differences, given the very small data set size. For evaluation on the final test set we therefore report results for all settings.

Results. Table 6 presents the results for error detection. The experiments with rule sets I and II use learned rule weights, the experiment with Set III uses hard constraints (i.e., rules with theoretically infinite weight).[13]

For **WSD**, we achieve the same results on all rule sets. In particular precision is in need of improvement. Compared to the other rule sets, Set II achieves less recall and precision. For **CR** the best overall performance (precision, recall and f-score) in error detection is achieved by

[13]Our experiments on Set III using learned weights could not be completed due to repeated process failures. In prior experiments we had obtained evaluation results close to Set I.

		P	R	F		P	R	F
$I_{d=0}$		34.8	100	51.6		67.5	82.6	74.3
$I_{d=1}$		34.8	100	51.6		**69.6**	**86.3**	**77.1**
$I_{d=2}$	WSD	34.8	100	51.6	CR	69.5	83.9	76.0
$I_{d=3}$		34.8	100	51.6				
II		34.1	96.8	50.4		68.5	82.6	74.9
III_{hard}		34.8	100	51.6		64.9	31.3	42.3

TABLE 6 Experiment results for error detection

using a fixed distance threshold of 1, i.e., by limiting the maximal distance between an entity centroid and the senses of its mentions to path length 1. Setting a higher threshold leads to a loss in both precision and recall. The latter figures look promising for automatic error detection as support for targeted annotation control. Set III yields lowest f-score for CR, which is mainly due to a very low recall. This indicates that the decision to reattach a mention to a new entity can not be based on distance alone. Instead, approaches using rule weights learned on data and thus tailored to distribution in the data achieve much better performance.

Overall, the figures in Table 6 show mixed results. For WSD, precision for error detection is devastatingly low. For CR, by contrast, we obtain very promising results of 69.6% precision at 77.1 points f-score that seem to reach a level of realistic cost-effectiveness to support manual annotation control.

Comparison of classical performance results for the sense and mention assignments predicted by the MLN inference rules in contrast to the original system assignments, however, show that automatic *error correction* is by far out of reach: The labeling performance of the predicted output of MLN inference drops by 5.06 (MUC)/6.41 (B^3) points f-score for CR and over 50 points f-score for WSD. Future work will investigate more refined constraint sets to obtain overall higher precision levels, in particular for WSD.

6 Conclusions and Future Work

To summarize, the contributions of our paper are two-fold: (i) We discussed performance issues in automatic semantic annotation of ritual texts and showed that domain adaptation can improve the annotation quality for WSD and CR. For frame-semantic annotation we could identify performance problems that can be addressed by retraining the semantic role labeling system on our semi-automatically annotated domain corpora, similarly to the domain adaptation methods employed

for preprocessing.

(ii) To further reduce the gap between automatic annotation quality and the high quality standards required for empirical research in Digital Humanities, we investigated a novel approach to error detection using Markov Logic as formal framework. Our approach to consistency control for semantic annotation explores inter-level dependencies between local (WSD) and discourse-level (CR) annotation decisions. Our experiments show promising results for detection of mistakes in automatic CR annotations, while error detection on sense assignment could not be achieved at a realistic level of performance.

In this paper, we could only come up with first investigations of this novel technique – with ample space for improvement.

First, our evaluation results are based on a small evaluation data set. Larger data sets are required to support statistically significant results and conclusions. Also, our current rule sets for consistency control rely on static and still noisy centroids computed on top of automatic CR and sense annotations. This severely restricts the induction of novel, more homogeneous discourse entities.

Our current experiments do not yet exploit the full power of Markov Logic Networks in that constraints for CR and WSD error detection are compiled in distinct rule sets. Future work will investigate joint processing of these constraints. We will also integrate the computation of centroids into the MLN inference process, so that changes in sense and mention assignment can more directly affect the computation of consistency constraints. Further improvements could be gained by including surface information for mentions and the formulation of constraints that implement the one-sense-per-discourse hypothesis. Finally, we will pursue deeper investigation of models similar to Set III that make error detection less dependent on optimization of distance thresholds.

Acknowledgments

This research has been funded by the German Research Foundation (DFG) and is part of the collaborative research center on ritual dynamics (Sonderforschungsbereich SFB-619, Ritualdynamik).[14]

We thank our student researchers Julio Cezar Rodrigues and Britta Zeller for assisting the experiments and providing gold standard linguistic annotations, as well as Borayin Maitreya Larios and Nils Jakob Liersch who provided frame-semantic annotations of the ritual texts. We further thank the anonymous reviewers for comments and suggestions.

[14] `http://www.ritualdynamik.de`

References

Agirre, Eneko and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pp. 33–41. Athens, Greece.

Bagga, Amit and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *Proceedings of the LREC 1998 Linguistic Coreference Workshop*, pp. 536–566. Granada, Spain.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 1, pp. 86–90.

Bögel, Thomas. 2011. *Entity-based Coreference Resolution combined with Discourse-New Detection*. Bachelor's thesis, Heidelberg University.

Chinchor, Nancy and Beth Sundheim. 2003. *Message Understanding Conference (MUC) 6*. Philadelphia: Linguistic Data Consortium.

Das, Dipanjan, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic Frame-Semantic Parsing. In *Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 948–956. L.A., California.

Daumé III, Hal. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263. Prague, Czech Republic.

Dickinson, Markus. 2010. Detecting Errors in Automatically-Parsed Dependency Relations. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden.

Dickinson, Markus and Chong Min Lee. 2008. Detecting Errors in Semantic Annotation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Marrakech, Morocco.

Dickinson, Markus and Detmar Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL)*. Budapest, Hungary.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities* 26(5–6):415–439.

Gutschow, Niels and Axel Michaels. 2005. *Handling Death. The Dynamics of Death and Ancestor Rituals Among the Newars of Bhaktapur*, vol. 3 of *Ethno-Indology. Heidelberg Studies in South Asian Rituals*. Harrassowitz Verlag.

Gutschow, Niels and Axel Michaels. 2008. *Growing Up. Hindu and Buddhist Initiation Rituals among Newar Children in Bhaktapur*, vol. 6 of *Ethno-Indology. Heidelberg Studies in South Asian Rituals*. Harrassowitz Verlag.

Hellwig, Oliver. 2011. *DCS - The Digital Corpus of Sanskrit*. Heidelberg.

Květoň, Pavel and Karel Oliva. 2002. (Semi-)Automatic Detection of Errors in PoS-Tagged Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (Coling)*. Taipei, Taiwan.

Loftsson, Hrafn. 2009. Correcting a POS-Tagged Corpus Using Three Complementary Methods. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pp. 523–531. Athens, Greece.

Navigli, Roberto and Mirella Lapata. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(4):678–692.

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

Reddy, Siva, Abhilash Inumella, Diana MacCarthy, and Mark Stevenson. 2010. IIITH: Domain Specific Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 387–391. Uppsala, Sweden.

Reiter, Nils, Oliver Hellwig, Anette Frank, Irina Gossmann, Borayin Maitreya Larios, Julio Rodrigues, and Britta Zeller. 2011. Adapting NLP Tools and Frame-Semantic Resources for the Semantic Analysis of Ritual Descriptions. In C. Sporleder, A. van den Bosch, and K. Z. Zervanou, eds., *Language Technology for Cultural Heritage, Foundations of Human Language Processing and Technology*. Springer.

Richardson, Matthew and Pedro Domingos. 2006. Markov Logic Networks. *Machine Learning* 62:107–136.

Soon, Wee Meng, Daniel Chung Yong Lim, and Hwee Tou Ng. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics* 27(4):521–544.

Versley, Yannick, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the ACL-08: HLT Demo Session*, pp. 9–12. Columbus, Ohio.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Conference on Message Understanding (MUC)*, pp. 45–52. Morristown, NJ, USA.

Yarowsky, David. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196. Cambridge, Massachusetts, USA.

Yu, Liang-Chih, Chung-Hsien Wu, and Eduard H. Hovy. 2008. OntoNotes: Corpus Cleanup of Mistaken Agreement Using Word Sense Disambiguation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling)*, pp. 1057–1064. Manchester, UK.