The structure of this paper is as follows: Section 2 describes the principal procedure of PANKOW. Section 3 describes the core algorithmic approach to categorizing instances from text. In Section 4, we present the empirical results of our evaluation. Then, we briefly discuss the integration into CREAM/OntoMat (Section 5). Before concluding the paper, we discuss related work in Section 6.

2. THE PROCESS OF PANKOW

This section gives a general overview of the process of PANKOW whereas Section 3 explains the concrete methods and Section 5 the implementation details. The process consists of four steps (depicted in Figure 1):

Input: A web page.

In our implementation, we assume that Web pages are handled individually in the CREAM/OntoMat framework ([18]), though actually batch processing of a whole Web site would be possible.

Step 1: The system scans the Web page for phrases in the HTML text that might be categorized as instances of the ontology. Candidate phrases are proper nouns, such as ‘Nelson Mandela’, ‘South Africa’, or ‘Victoria Falls’. We use a part-of-speech tagger (cf. Section 3 and Section 5) to find such candidate proper nouns.

Thus, we end up with a

Result 1: set of candidate proper nouns

Step 2: The system iterates through the candidate proper nouns. It uses the approach described in Section 3.1, introducing all candidate proper nouns and all candidate ontology concepts into linguistic patterns to derive hypothesis phrases. For instance, the candidate proper noun ‘South Africa’ and the concepts Country and Hotel are composed into a pattern resulting in hypothesis phrases like ‘South Africa is a country’ and ‘South Africa is a hotel’.

Result 2: Set of hypothesis phrases.

Step 3: Then, Google™ is queried for the hypothesis phrases through its Web service API (Section 3.2). The API delivers as its results

Result 3: the number of hits for each hypothesis phrase.

Step 4: The system sums up the query results to a total for each instance-concept pair. Then the system categorizes the candidate proper nouns into their highest ranked concepts (cf. Section 3.3). Hence, it annotates a piece of text as describing an instance of that concept. Thus we have

Result 4: an ontologically annotated web page.

In principle, the query results of step 3 could be investigated further. For instance, it could make sense to constrain the number of hits for hypothesis phrases to the ones that occur in Web pages with topics closely related to the topic of the current Web page, as, e.g. measured in terms of cosine distance of the documents. However, without direct access to the Google™ databases we have considered this step too inefficient for use in automatic annotation and hence ignore it in the following.

3. PATTERN-BASED CATEGORIZATION OF CANDIDATE PROPER NOUNS

There is some history of applying linguistic patterns to identify ontological relationships between entities referred to in a text. For instance, Heairst([19]) as well as Charniak and Berland [6] make use of such a pattern-based approach to discover taxonomic and part-of relations from text, respectively. Hahn and Schnattinger [15] also make use of such patterns and incrementally established background knowledge to predict the correct ontological class for unknown named entities appearing in a text. The core idea of any such pattern-based approach is that one may justify an ontological relationship with reasonable accuracy when one recognizes some specific idiomatic/syntactic/semantic relationships. Grammar to the pattern-based approach is that the specifically addressed idiomatic/syntactic/semantic relationships may be very easily spotted because they may be typically specified through simple and efficiently processable regular expressions.

In the following we first present the set of patterns, in a second step we describe the procedure to actually search for them and finally we explain how we use the information conveyed by them for the actual classification of instances.

3.1 Patterns for Generating Hypothesis Phrases

In the following we describe the patterns we exploit and give a corresponding example from the data set that we used for empirical evaluation (cf. Section 4).

3.1.1 Hearst Patterns

The first four patterns have been used by Hearst to identify isa-relationships between the concepts referred to by two terms in the text. However, they can also be used to categorize a candidate proper noun into an ontology.

Since the entities denoted by candidate proper nouns are typically modeled as instances of an ontology, we also describe the problem as conveniently as the instantiation of a concept from a given ontology. Correspondingly, we formulate our patterns using the variable '<CONCEPT>' to refer to a candidate noun phrase, as the name of an ontology instance, and '<CONCEPT>' to refer to the name of a concept from the given ontology.

The patterns reused from Hearst are:

H1: <CONCEPT>s such as <INSTANCE>
H2: such <CONCEPT>s as <INSTANCE>
H3: <CONCEPT>s, (especially including) <INSTANCE>
H4: <INSTANCE> (and/or) other <CONCEPT>s

The above patterns would match the following expressions (in this order): hotels such as Ritz; such hotels as Hilton; presidents, especially George Washington; and the Eiffel Tower and other sights in Paris.

3.1.2 Definites

The next patterns are about definites, i.e. noun phrases introduced by the definite determiner 'the'. Frequently, definites actually refer to some entity previously mentioned in the text. In this sense, a phrase like 'the hotel' does not stand for itself, but it points as a so-called anaphora to a unique hotel occurring in the preceding text. Nevertheless, it has also been shown that in common texts
Towards the Self-Annotating Web

Philipp Cimiano, Siegfried Handschuh, Steffen Staab
1Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
2Ontoprise GmbH, 76131 Karlsruhe, Germany

ABSTRACT
The success of the Semantic Web depends on the availability of ontologies as well as on the proliferation of web pages annotated with metadata conforming to these ontologies. Thus, a crucial question is where to acquire these metadata. In this paper we propose PANKOW (Pattern-based Annotation through Knowledge on the Web), a method which employs an unsupervised, pattern-based approach to categorize instances with regard to an ontology. The approach is evaluated against the manual annotations of two human subjects. The approach is implemented in OntoMat, an annotation tool for the Semantic Web and shows very promising results.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—Indexing Methods; I.7.1 [Document and Text Processing]; I.2.7 [Natural Language Processing]

General Terms
Measurement, Documentation, Design, Experimentation, Human Factors, Languages

Keywords
Semantic Annotation, Metadata, Information Extraction, Semantic Web

1. INTRODUCTION

The Semantic Web builds on contents that are described semantically via ontologies and metadata conforming to these ontologies. While one sees a multitude of ontologies and ontology-like structures being defined in research projects like DAML, in standardization efforts like the ones by OASIS and in industrial endeavors like UDDI, corresponding metadata are mostly missing.

The reason is that in spite of methods and tools like large-scale information extraction [9], learning of information extraction rules [7], and the application of both in current annotation toolsets [17, 28], the obstacles for producing such markup remain high:

• Manual definition of an information extraction system is a laborious task requiring a lot of time and expert know-how [13]; and
• Learning of extraction rules requires a lot of, frequently too many, examples for learning the rules.

Here, one encounters a vicious circle where there is no Semantic Web because of a lack of metadata, and there are no metadata because there is no Semantic Web that one could learn from, for example, by training an IE system such as Amilcare [7].

As a way out of this vicious cycle we propose a new paradigm: the Self-annotating Web. The principal idea of the self-annotating Web is that it uses globally available Web data and structures to semantically annotate—or at least facilitate annotation of—local resources. Initial blueprints for this paradigm are found in such works as the following:

• Some researchers use explicit, linguistically motivated natural language descriptions to propose semantic relationships [6, 13, 19, 22].

• Others use the Web to cope with data sparseness problems in tasks that require statistics about possible semantic relationships [1, 14, 21, 25].

• In [10, 12], the Web structure itself is used to determine a focus for harvesting data. Thus, specialized semantic relationships, such as recommendations coming from a particular Web community, can be derived.

Going a step towards the Semantic Web, we propose an original method called PANKOW (Pattern-based Annotation through Knowledge On the Web), which employs an unsupervised, pattern-based approach to categorize instances with regard to a given ontology.

The approach is novel, combining the idea of using linguistic patterns to identify certain ontological relations as well as the idea of using the Web as a big corpus to overcome data sparseness problems. It is unsupervised as it does not rely on any training data annotated by hand and it is pattern-based in the sense that it makes use of linguistically motivated regular expressions to identify instance-concept relations in text. The driving principle behind PANKOW is one of disambiguation by maximal evidence in the sense that for a given instance it proposes the concept with the maximal evidence derived from Web statistics. The approach thus bootstraps semantic annotations as it queries the Web for relevant explicit natural-language descriptions of appropriate ontological relations.

PANKOW has been conceived for our annotation framework CREAM [16] and has been implemented in OntoMat2 using queries to the Web service API of GoogleTM. The automatic annotation produced by PANKOW has been evaluated against semantic annotations produced by two independent human subjects.