Learning by Googling

Philipp Cimiano
Institute AIFB
University of Karlsruhe
cimiano@aifb.uni-karlsruhe.de

Steffen Staab
Institute for Computer Science
University of Koblenz-Landau
staab@uni-koblenz.de

ABSTRACT

The goal of giving a well-defined meaning to information is currently shared by endeavors such as the Semantic Web as well as by current trends within Knowledge Management. They all depend on the large-scale formalization of knowledge and on the availability of formal metadata about information resources. However, the question how to provide the necessary formal metadata in an effective and efficient way is still not solved to a satisfactory extent. Certainly, the most effective way to provide such metadata as well as formalized knowledge is to let humans encode them directly into the system, but this is neither efficient nor feasible. Furthermore, as current social studies show, individual knowledge is often less powerful than the collective knowledge of a certain community.

As a potential way out of the *knowledge acquisition bottleneck*, we present a novel methodology that acquires collective knowledge from the World Wide Web using the Google^{TM} API. In particular, we present PANKOW, a concrete instantiation of this methodology which is evaluated in two experiments: one with the aim of classifying novel instances with regard to an existing ontology and one with the aim of learning sub-/superconcept relations.

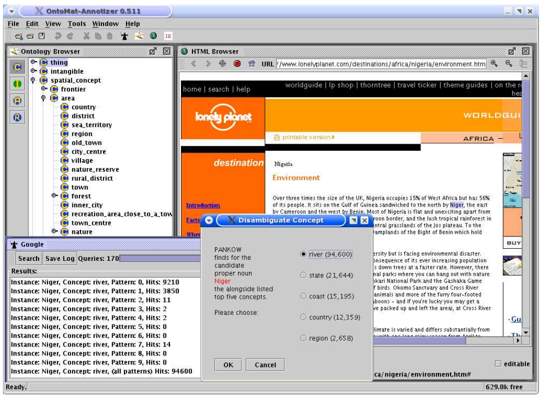
1. INTRODUCTION

The goal of giving a well-defined meaning to information is currently shared by different research communities. This goal is based on the assumption that, once information has a well-defined meaning, it can be (i) searched and retrieved more effectively, (ii) shared between different parties and (iii) used to derive implicit or new knowledge via certain inference mechanisms. This vision is shared in particular by the Semantic Web [5], by current trends within Knowledge Management [22] as well as by knowledge-based information systems in general. As has been argued in almost every work dealing with knowledge acquisition, any information system relying on background knowledge suffers from the so called *knowledge acquisition bottleneck*, i.e. the difficulty of encoding knowledge into a system in a declarative fashion. So far, it seems that the most effective – though certainly not most efficient – way of dealing with this problem is to let a group of knowledge engineers model the required world knowledge from scratch.

Inspired by current social studies as [43], in which it is argued that collective knowledge is much more powerful than individual knowledge, we present in this paper a new paradigm of dealing with the above mentioned bottleneck. In very general terms our paradigm is based on the idea that collective knowledge is gathered as a first step and then as a second step presented to a knowledge engineer who can thus effectively and efficiently customize this collective knowledge with regard to the specific context of interest. In this model, the purpose of general knowledge is to compensate the potential lack of knowledge of an individual with respect to a certain topic, while the role of the individual is to filter the collective knowledge with regard to a specific context.

This abstract model with the purpose of overcoming the knowledge acquisition bottleneck is for example instantiated by our PANKOW (Pattern-based Annotation through Knowledge on the Web) methodology [13]. PANKOW was originally conceived to support a web-page annotator in the task of assigning the instances appearing in the page to the appropriate concept in a given ontology in line with the CREAM framework [28]. In particular, PANKOW generates instances of lexico-syntactic patterns indicating a certain semantic or ontological relation and counts their occurrences in the World Wide Web using the Google^{TM}API. The statistical distribution of instances of these patterns then constitutes the collective knowledge which is taken into account by the annotator to decide with which concept to annotate the instance in the particular context. Figure 1 for example shows a dialog in which the user is presented with the top 5 suggestions from the collective knowledge about how to annotate the instance *Niger*, i.e. as a river, as a country, etc. The advantage of such an approach combining collective and individual knowledge to overcome the knowledge-acquisition bottleneck seems thus obvious: even if the individual has never heard about the instance in question, together with the collective knowledge and the local context in which the instance appears, he might get a fairly accurate idea of the concept it belongs to.

The remainder of this article is further structured as follows: in Section 2 we describe PANKOW in more detail and present our lexico-syntactic pattern library. We also discuss the application of PANKOW in the annotation scenario described above as well as to learning sub-/superconcept relations. In section 3 we present results of an evaluation of PANKOW with respect to both tasks. Finally, before concluding, we discuss some related work.



URL '/www.lonelyplanet.com/destinations/africa/nigeria/environment.htm

Nigeria

Environment

Over three times the size of the UK, Nigeria occupies 15% of West Africa but has 56% of its people. It sits on the Gulf of Guinea sandwiched to the north by Niger, the east by Cameroon and the west by Benin. Most of Nigeria is flat and unexciting apart from
roon border, and the lush tropical rainforest in
ntral grasslands of the Jos plateau. To the
amplands of the Bight of Benin which hold

ersity but is facing environmental disaster.
onsequence of its ever increasing population
s down trees at a faster rate. However, there
nal parks where you can hang out with nature
nkari National Park and the Gashaka Game
f birds. Okomo Sanctuary and Cross River
animals and more of the furry four-footed
boons - and if you're lucky you may get a
ve packed up and left the area), at Cross River

limate is varied and differs substantially from

ca/nigeria/environment.htm#

Queries: 170

Results:

```
Instance: Niger, Concept: river, Pattern: 0, Hits: 9210
Instance: Niger, Concept: river, Pattern: 1, Hits: 3850
Instance: Niger, Concept: river, Pattern: 2, Hits: 11
Instance: Niger, Concept: river, Pattern: 3, Hits: 2
Instance: Niger, Concept: river, Pattern: 4, Hits: 2
Instance: Niger, Concept: river, Pattern: 5, Hits: 0
Instance: Niger, Concept: river, Pattern: 6, Hits: 0
Instance: Niger, Concept: river, Pattern: 7, Hits: 14
Instance: Niger, Concept: river, Pattern: 8, Hits: 0
Instance: Niger, Concept: river, Pattern: 9, Hits: 0
Instance: Niger, Concept: river, (all patterns) Hits: 94600
```

Disambiguate Concept

PANKOW finds for the candidate proper noun Niger the alongside listed top five concepts.

Please choose:

● river (94,600)
○ state (21,644)
○ coast (15,195)
○ country (12,359)
○ region (2,658)

OK Cancel

Figure 1: PANKOW within an annotation scenario (interactive mode)

2. PANKOW

PANKOW is based on the idea that certain lexico-syntactic patterns matched in texts convey a specific semantic relation. Pioneering research in this line was conducted by Hearst [29] who defined a collection of patterns indicating sub-/superconcept relations. An example of such a pattern used by Hearst is the following:

$$\text{such } NP_0 \text{ as } NP_1,...,NP_{n-1} \text{ (or|and) other } NP_n$$

where NP stands for a noun phrase. If such a pattern is matched in a text, according to Hearst we could derive that for all $0 < i \leq n$ hyponym(lemma(NP_i),$lemma(NP_0)$)[1], where lemma(NP) represents the lemma[2] of the concatenation of each open-class[3] word in NP.[4] For example, from the sentence '*Such injuries as bruises, wounds and broken bones...*' we could derive the relations: hyponym(bruise,injury), hyponym(wound,injury) and hyponym(broken bone,injury). Moreover, PANKOW also builds upon the idea that such patterns as described above can not only be matched in a corpus, but also in the World Wide Web as in [13], [17], [18] or [36].

For this purpose, PANKOW generates pattern instances out of pattern schemes and counts the hits of these pattern instances on the web. For each instance or concept of interest, we thus yield the number of times it is related to other entities in the specific way indicated by the pattern schema, thus yielding a statistical 'fingerprint' for this object with respect to a given semantic relation. In what follows, we first describe the process from a general point of view. Then, in Section 2.2 we describe the patterns we use and finally we formally define what a statistical fingerprint is and how it can be used.

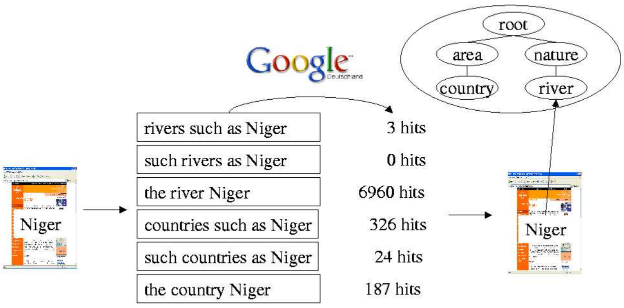


Figure 2: PANKOW within an annotation scenario (automatic mode)

2.1 The Process of PANKOW

In this paper we slightly abstract from the process of PANKOW as described in [13]. In fact, the general process consists of three steps:

Input: a set of entities (instances or concepts) to be classified with regard to an ontology

Step 1: The system iterates through the set of entities to be classified and generates instances of patterns, one for each concept in the ontology. For example, the instance '*South Africa*' and the concepts Country and Hotel are composed using a pattern schema of our pattern library (see 2.2) and resulting in pattern instances like '*South Africa is a country*' and '*South Africa is a hotel*' or '*countries such as South Africa*' and '*hotels such as South Africa*'

Result 1: Set of pattern instances

Step 2: Then, Google^{TM} is queried for the pattern instances through its Web service API. The API delivers as its results

Result 2: the counts for each pattern instance

Step 3: The system sums up the query results to a total for each concept.

Result: The statistical fingerprint for each entity, i.e. the results of aggregating for each entity the number of Google counts for all pattern instances conveying the relation of interest

The statistical fingerprint then represents the collective knowledge about the potential concepts an instance could belong to or about the potential superconcepts of a certain concept. Given the tasks of (i) classifying instances with regard to an ontology or (ii) finding an appropriate superconcept for a new concept, a knowledge engineer could be presented with the most relevant *view* of a statistical fingerprint in order to take a final decision.

Figure 2 depicts an example of how PANKOW can be employed in an annotation scenario. An important question here is how to find potential new instances in web pages. Though this is not directly the topic of this paper, a few

[1] From a linguistic point of view, a term t_1 is a hyponym of a term t_2 if we can say '*a* t_1 *is a* t_2'. Correspondingly, t_2 is then a hypernym of t_1.

[2] The lemma of a word is its base or normal form, i.e. *cats* - *cat*, *drove* - *drive*, etc.

[3] In contrast to closed-class words which belong to a class of words with a constant extension (examples are prepositions, determiners, ...), open-classes are evolving classes whose extension constantly changes.

[4] Hearst doesn't explicitly talk about lemmatization, but it is clear from her examples that lemmatization should be performed.

words on this issue seem appropriate. In order to find candidate new instances or concepts in a web page to be annotated, we first extract the textual content of the web pages and then run a part-of-speech tagger[5] over the page to assign each token its corresponding syntactic category. Then certain regular expressions defined over these tags and the corresponding tokens allow to find candidate instances or concepts. The main heuristic in finding instances consists of finding sequences of capitalized words tagged as proper nouns. Thus, in the abstract of this paper our method would find *Semantic Web*, *Knowledge Management*, *World Wide Web*, *Google API* and *PANKOW* as potential instances to be annotated. In order to find concepts in web pages, we interpret each sequence of lower case words tagged as common nouns as potential concepts. Of course, the patterns we use exploit other heuristics, but a detailed description of these is out of the scope of this paper and in general the issue of how to find candidate instances or concepts is orthogonal to the aim of the approach described in this paper.

Now given an unknown instance or concept on a certain web page, patterns respectively indicating an *instance-of* or *subconcept* relation are instantiated for the new instance or concept and each concept in the target ontology. Finally, given the statistical fingerprint of the instance or concept, we follow a principle of disambiguation by maximal evidence thus assigning the instance or concept to that concept in the target ontology with the highest number of hits in the statistical fingerprint. Figure 2 illustrates the principle of disambiguation by maximal evidence within an annotation scenario. Instead, a user can also be involved in the process and for example asked to select a concept out of the top 5 elements of the statistical fingerprint. Figure 1 depicts an annotation scenario in which the user is asked to choose one concept out of the top-5 view of the statistical fingerprint.

2.2 The Pattern Library

In the following we describe the patterns we exploit and give a corresponding example.

2.2.1 Hearst Patterns

The first four patterns have been used by Hearst to identify *isa*-relationships between the concepts referred by two words in the text. However, they can also be used to spot *instance-of*-relations. In fact, in PANKOW they are used as indicating *subclass* as well as *instance-of* relations, depending on whether the entity to be classified is an instance or a concept. Correspondingly, we formulate our patterns using the variable '<I>' to refer to the name of an instance and the variable '<C>' to refer to the name of a concept from the given ontology.

The patterns reused from Hearst are:

HEARST1: $< C >$s such as $< I|C' >$

HEARST2: such $< C >$s as $< I|C' >$

HEARST3: $< C >$s, (especially|including) $< I|C' >$

HEARST4: $< I|C' >$ (and|or) other $< C >$s

Depending on whether we are attempting to classify an instance or a concept, we would then either derive: instance-of(I,C) or subconcept(C',C). The above patterns would match the following expressions:

continents such as Asia (HEARST1)
vehicles such as cars (HEARST1)
such continents as Africa (HEARST2)
such cars as cabriolets (HEARST2)
presidents, especially George Washington (HEARST3)
vehicles, especially motor-bikes (HEARST3)
the Eiffel Tower and other sights in Paris (HEARST4)
motor-bikes and other two-wheeled vehicles (HEARST4)

2.2.2 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 more than 50% of all definite expressions are *non-referring* [38], i.e. they exhibit sufficient descriptive content to enable the reader to uniquely determine the entity referred to from the global context. For example, the definite description '*the Hilton hotel*' has sufficient descriptive power to uniquely pick-out the corresponding real-world entity for most readers. One may deduce that '*Hilton*' is the name of the real-world entity of type Hotel to which the above expression refers.

Consequently, we apply the following two patterns to categorize an instance by definite expressions:

DEFINITE1: the $< I > < C >$

DEFINITE2: the $< C > < I >$

The first and the second pattern would for example match the expressions '*the Hilton hotel*' and '*the hotel Hilton*', respectively. It is important to mention that these patterns are in our approach only used to categorize instances into the ontology, but not concepts.

2.2.3 Apposition and Copula

The following pattern makes use of the fact that certain entities appearing in a text are further described in terms of an apposition as in '*Excelsior, a hotel in the centre of Nancy*'. The pattern capturing this intuition looks as follows:

APPOSITION: $< I|C' >$, a $< C >$

The probably most explicit way of expressing that a certain entity is an instance or a subconcept of a certain concept is by the verb '*to be*' in a copula[6] construction as for example in '*The Excelsior is a nice hotel in the center of Nancy*'. Here's the general pattern:

COPULA: $< I|C' >$ is a $< C >$

[5] A part-of-speech tagger assigns syntactic categories to words. We use the QTag tagger in http://web.bham.ac.uk/o.mason/software/tagger/.

[6] A copula is an intransitive verb which links a subject to an object, an adjective or a constituent denoting a property of the subject.

2.3 Statistical Fingerprints

Having defined these patterns, one could match these patterns in a corpus and propose the corresponding relations. However, it is well known that the above patterns are rare and thus one will need a sufficiently big corpus to find a significant number of matches.

Thus, PANKOW resorts to the biggest corpus available: the World Wide Web. In fact, several researchers have shown that using the Web as a corpus is an effective way of addressing the typical data sparseness problem one encounters when working with corpora (compare [26], [32], [36], [40]). Actually, we subscribe to the principal idea by Markert *et al.* [36] of exploiting the Google^{TM} API. As in their approach, rather than actually downloading web pages for further processing, we just take the number of web pages in which a certain pattern appears as an indicator for the strength of the pattern.

Given a candidate entity we want to classify with regard to an existing ontology, we instantiate the above patterns with each concept from the given ontology. For each pattern instance, we query the Google^{TM} API for the number of documents that contain it. The function 'count' models this query.

$$\text{count} : E \times C \times P \rightarrow \mathbb{N}$$

Thereby, E, C and P stand for the set of all entities to be classified, for the concepts from a given ontology and for a set of pattern schema, respectively. Thus, $\text{count}(e, c, p)$ returns the number of hits of pattern the pattern schema p instantiated with the entity e and the concept c. Further we define the sum over all the patterns conveying a certain relation r:

$$count_r(e, c) = \sum_{p \in P_r} \text{count}(e, c, p)$$

where P_r is the set of pattern schemes denoting a certain relation r.

Now we formally define the statistical fingerprint of an entity e with respect to a relation r and a set of concepts C:

$$SF(e, r, C) := \{(c, n) |\ c \in C \wedge n = count_r(e, c)\}$$

Further, instead of considering the complete statistical fingerprints, we consider views of these such as defined by the following formulas. The first formula defines a view of the statistical fingerprint which only contains the concept with maximal number of hits.[7]

$$\begin{aligned} SF_{max}(e, r, C) := \{(c, n) |\ & c := argmax_{c' \in C} count_r(e, c') \wedge \\ & n = count_r(e, c)\} \end{aligned}$$

Further, we extend this to consider the top-m concepts with maximal count:

$$\begin{aligned} SF_m(e, r, C) := \{(c, n) |\ & C = \{c_1, c_2, ..., c_{|C|}\} \wedge \\ & count_r(e, c_1) \leq ... \leq count_r(e, c_{|C|}) \wedge \\ & c \in \{c_1, ..., c_m\} \wedge n = count_r(e, c)\} \end{aligned}$$

if $m \leq |C|$.

Finally, we also consider a view only taking into account those concepts having hits over a certain threshold θ:

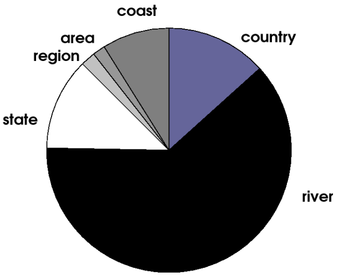


Figure 3: View of Niger's Fingerprint

$$SF_\theta(e, r, C) := \{(c, n) |\ count_r(e, c) \geq \theta \wedge n = count_r(e, c)\}$$

We can now combine these views by set operations. For example, we yield the set of the m top concepts having hits over a threshold θ as follows:

$$SF_{m,\theta}(e, r, C) = SF_m(e, r, C) \cap SF_\theta(e, r, C)$$

As an example of such a view, consider the visualization of the SF_6 view of the statistical fingerprint for Niger with regard to the *instance-of* relation in Figure 3. It is interesting to observe that the most prominent concept for *Niger* seems to be *river*, directly followed by *country* and further by *state, coast, region* and *area.*

3. EVALUATION

We have evaluated PANKOW with respect to two tasks: the task of finding the appropriate ontological concept for a given instance, and the task of finding sub/superconcept relations.

3.1 Instance Classification Experiment

For our instance classification experiment, we asked 2 subjects to annotate 30 texts with destination descriptions from *http://www.lonelyplanet.com/destinations.* They used a pruned version of the tourism ontology developed within the GETESS project [42]. We manually pruned this ontology by removing concepts which did not appear in the above pages in order to facilitate the annotation process. The original ontology consisted of 1043 concepts, while the pruned one consisted of 682. The subjects were told to annotate instances in the web page with the appropriate concept from the ontology. In what follows, we will refer to these subjects as A and B. Subject A actually produced 436 categorizations and subject B produced 392. There were 277 proper nouns (referred to by I in the following; $|I| = 277$) that were annotated by both subjects. For these 277 proper nouns, they used 59 different concepts (henceforth constituting our set of concepts C). The categorial agreement on these 277 proper nouns as measured by the Kappa statistic (cf. [11]) was 63.48%, which allows to conclude that the classification task is overall well defined. In the following, we only consider the common instances in I for our evaluation.

3.1.1 Evaluation Measures

To evaluate our approach, we compare the answers of our system with the following reference standards:

[7] We assume that argmax breaks ties randomly in this context.

- $Standard_A := \{(i,c) |$ for each $i \in I$ the categorization $c \in C$ produced by subject A$\}$
- $Standard_B := \{(i,c) |$ for each $i \in I$ the categorization $c \in C$ produced by subject B$\}$

Now as answers $S_{max,\theta}$ of the system we consider the following set:

$$S_{max,\theta} := \{(i,c) | i \in I \wedge \{(c,n)\} = SF_{max,\theta}(i, instance{-}of, C)\}$$

As evaluation measures, we use the well-known P(recision), R(ecall) and F_1-Measures to evaluate our system against $Standard_A$ and $Standard_B$. P, R and F_1 are defined as follows (for $y \in \{A, B\}$, the two standards):

$$P_y = \frac{|\text{correct answers}|}{|\text{total answers}|} = \frac{|S_{max,\theta} \cap Standard_y|}{|S_{max,\theta}|}$$

$$R_y = \frac{|\text{correct answers}|}{|\text{answers in reference standard}|} = \frac{|S_{max,\theta} \cap Standard_y|}{|I|}$$

$$F_{1,y} = \frac{2 * P_y * R_y}{P_y + R_y}$$

Furthermore, in our experiments we will always average the results for both annotators as given by the following formulas:

$$P_{avg} = \frac{P_A + P_B}{2}$$

$$R_{avg} = \frac{R_A + R_B}{2}$$

$$F_{1,avg} = \frac{F_{1,A} + F_{1,B}}{2}$$

To get an upper bound for the task we are looking at, we also calculated the F_1-Measure of $Standard_A$ measured against $Standard_B$ and the other way round and got F_1=62.09% as average. This value thus represents an upper bound for any system attempting to find the correct class for an unknown instance.

3.1.2 Results of Instance Classification Experiment

Table 1 shows the top 60 $SF_{max}(i,\text{instance-of}, C)$ values for different instances i. While some classifications are definitely spurious, it can be seen in general that the results are quite reasonable. Figure 4 shows the precision, recall and F_1-Measure values for different thresholds θ within the interval [0..1000], averaged over both reference standards: $Standard_A$ and $Standard_B$. Obviously, the precision increases roughly proportionally to the threshold θ, while the recall and F_1-Measure values decrease. It can be observed that P=R=F at $\theta = 0$. The best $F_{1,avg}$-Measure was 28.24% at a threshold of $\theta = 60$ and the best Recall (R_{avg}) was 24.9% at a threshold of $\theta = 0$.

In a second version of the experiment, instead of merely choosing the concept with maximal count with respect to the statistical fingerprint, we considered the top 5 concepts, i.e. the view $SF_{5,\theta} = SF_5 \cap SF_\theta$ and considered the answer

Instance	Concept	# Google^{TM} Matches
Atlantic	city	1520837
Bahamas	island	649166
USA	country	582275
Connecticut	state	302814
Caribbean	sea	227279
Mediterranean	sea	212284
South Africa	town	178146
Canada	country	176783
Guatemala	city	174439
Africa	region	131063
Australia	country	128607
France	country	125863
Germany	country	124421
Easter	island	96585
St Lawrence	river	65095
Commonwealth	state	49692
New Zealand	island	40711
Adriatic	sea	39726
Netherlands	country	37926
St John	church	34021
Belgium	country	33847
San Juan	island	31994
Mayotte	island	31540
EU	country	28035
UNESCO	organization	27739
Austria	group	24266
Greece	island	23021
Malawi	lake	21081
Israel	country	19732
Perth	street	17880
Luxembourg	city	16393
Nigeria	state	15650
St Croix	river	14952
Nakuru	lake	14840
Kenya	country	14382
Benin	city	14126
Cape Town	city	13768
St Thomas	church	13554
Niger	river	13091
Christmas Day	day	12088
Ghana	country	10398
Crete	island	9902
Antarctic	continent	9270
Zimbabwe	country	9224
Central America	region	8863
Reykjavik	island	8381
Greenland	sea	8043
Cow	town	7964
Expo	area	7481
Ibiza	island	6788
Albania	country	6327
Honduras	country	6143
Iceland	country	6135
Nicaragua	country	5801
Yugoslavia	country	5677
El Salvador	country	5154
Senegal	river	5139
Mallorca	island	4859
Nairobi	city	4725
Cameroon	country	4611
Rust	park	4541

Table 1: Top 60 Instance-Concept Relations

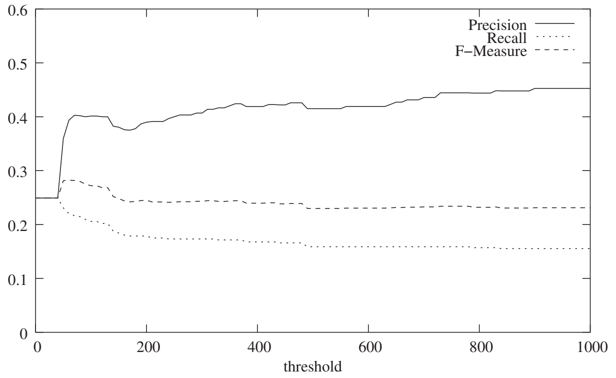


Figure 4: Precision, Recall and F_1-Measure for $S_{max,\theta}$ over threshold θ (instance classification)

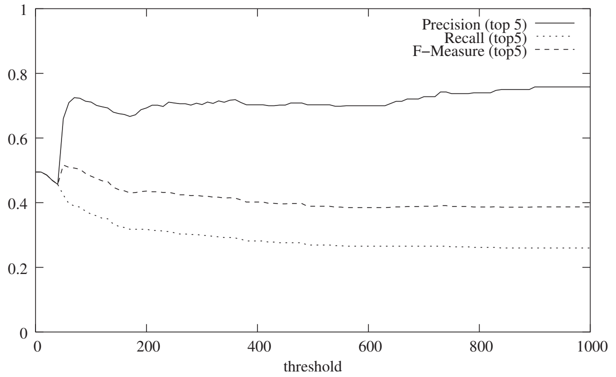


Figure 5: Precision, Recall and F_1-Measure and Recall for $S_{5,\theta}$ over threshold θ (instance classification)

as correct if the annotator's answer was in this view. The results in terms of the same measures are given in figure 5. The qualitative behaviour of the three measures is similar as in the first experiment, but obviously the results are much better. The best F_1-Measure of 51.64% was reached at a threshold of $\theta = 50$, corresponding to a Precision of 66.01% and a recall of 42.42%. Concluding, these results mean that in 66% of the cases the correct concept for an instance is among the top 5 suggestions and on the other hand for more than 40% of the relevant instances the system is able to suggest 5 concepts, one of which is the correct one. This is certainly a very satisfactory result and a good proof that using our PANKOW methodology to gather collective knowledge in form of statistical fingerprints and presenting certain views of these to a user would drastically help to reduce the time taken to annotate a given web page.

3.2 Sub-/Superconcepts Extraction

As a second experiment, we attempted to reproduce the sub-/superconcept relations of a given ontology. In particular, we considered the tourism ontology which was manually constructed by an ontology engineer in the context of the comparison study described in [35]. Furthermore, as this ontology was specified in German, we translated it into English. The ontology consisted of 289 concepts, from which we removed a few abstract concepts such as *partially_material_thing*, or *geometric_concept* thus yielding 272 concepts with 225 direct is-a relations and 636 transitive (direct + non-direct) is-a relations between them. For our evaluation we take into account the set of transitive relations.

3.2.1 Evaluation Measures

As in the first experiment, we evaluated PANKOW in terms of Precision, Recall and F_1-Measure. In contrast to the above experiment, we merely compared to one reference standard, i.e. the ontology described above. The answers of the system are now defined as follows:

$$S_{max,\theta} := \{(c', c) |\ \{(c, n)\} = (SF_{max,\theta}(c', subconcept, C)\}$$

The reference standard is given by the following set O:

$$O := \{(c', c) |\ c' \leq_C c\}$$

where $\leq_C$ is the partial order representing the concept hierarchy of the reference ontology.
Now, Precision, Recall and F_1-Measure are defined as follows:

$$P = \frac{|S_{max,\theta} \cap O|}{|S_{max,\theta}|}$$

$$R = \frac{|S_{max,\theta} \cap O|}{|O|}$$

$$F_1 = \frac{2 * P * R}{P + R}$$

3.2.2 Results of Sub-/Superconcept Extraction

Figure 6 shows the results of the sub-/superconcept extraction in terms of Precision, Recall and F_1-Measure. In this case the best F_1-Measure was F_1=18.25% and was reached at threshold $\theta = 0$, corresponding to a precision of P=21.74% and a recall of R=15.73%. This was also the overall best recall. Thus, the results seem to be not as good as in the above experiment. This is probably due to the fact that concept labels are much more ambiguous than instance labels. When considering again the top 5 best suggestions of the system, the results increase as shown in figure 7. The best F_1-Measure in this second version of the experiment was F=52.33% at $\theta = 0$; the precision was P=62.32% and the recall R=45.10%. These results are also impressive and again corroborate the claim that our approach is a very promising step towards overcoming the knowledge acquisition bottleneck.

3.3 Discussion

Our experiments have shown that the results of our system are within a range in which they can not be used automatically without any human interaction. However, we have also shown that when operating in an interactive mode in which a user is presented with the top 5 suggestions, our system performs very well obtaining F-Measures over 50% on such non-trivial classification tasks. However, we compare our approach from a quantitative point of view with systems performing the task of assigning instances to the corresponding concept automatically. In the computational linguistics community this task is known as 'Named Entity Recognition and Classification' (NERC). This task received special

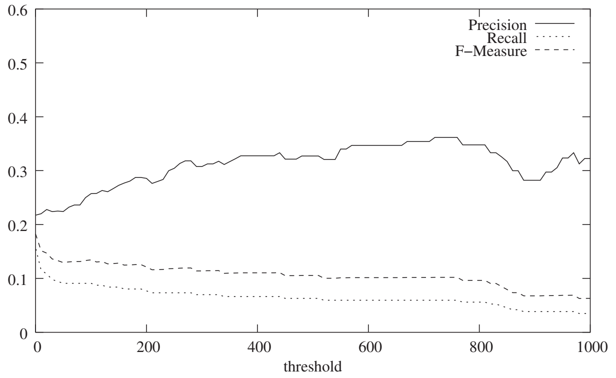


Figure 6: Precision, Recall and F_1-Measure for $S_{max,\theta}$ over threshold θ (sub-/superconcept extraction)

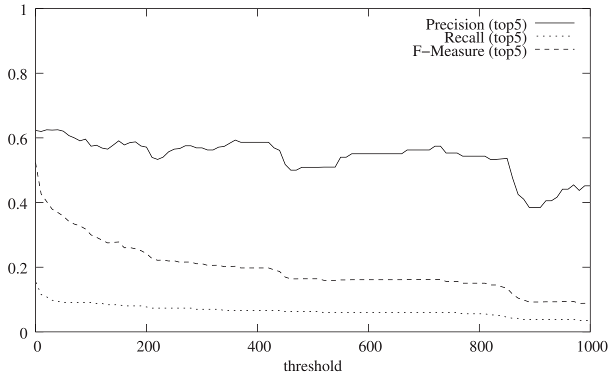


Figure 7: Precision, Recall and F_1-Measure for $S_{5,\theta}$ over threshold θ (sub-/superconcept extraction)

attention as a subtask within the framework of the Message Understanding Conferences (MUC) ([30]) which aimed at evaluating information extraction systems on a shared task. The named entity recognition task comprised three categories: PERSON, LOCATION and ORGANIZATION and systems typically achieved F-Measures well above 90%. However, this task is certainly much simpler than the ones in [3] where 1200 WordNet synsets are considered, [27] which consider 325 concepts, [23] taking into account 8 and [20] considering from 2-8 depending on the document in question. These systems are described in detail in the following section 4. Table 2 gives an overview of these systems, in particular showing the number of classes considered, the type of text preprocessing needed as well as the recall or accuracy on the task. It is important to mention here that as in our case the set of instances annotated by the system is equal to the instances annotated by the human subject, the recall corresponds to the accuracy results reported by the other systems. It can be concluded from the table that the performance of our system, given the number of classes considered and the fact that no text processing methods are needed, seems indeed reasonable compared to systems performing a related task. It is however important to emphasize that as the number of classes considered is not the same, the systems are not directly comparable.

4. RELATED WORK

Traditionally, supervised information extraction techniques have been applied to facilitate the creation of metadata on the basis of textual input. Several learning techniques have been applied to induce extraction rules from a labeled set of training examples. Kushmerick et al. for example developed a technique called Boosted Wrapper Induction [24]. Califf and Mooney [9] use ILP-based bottom-up rule induction techniques, while Soderland [41] uses a top-down rule induction algorithm applying a hill-climbing approach. Recently, Ciravegna [15] developed a novel algorithm called LP^2. However, due to the fact that all these systems exploit regularities in the induction of extraction rules, their application for information extraction from the Web seems limited. Furthermore, the cost for using such systems remains extremely high as one needs to provide a considerable amount of training examples.

Concerning the task of learning the correct class or ontological concept for an unknown entity, there was quite a lot of related work within the framework of the above mentioned Message Understanding Conferences. However, the challenge of categorizing into 3 classes is quite modest when compared against the challenge of categorizing into 59 classes as in our approach. We thus focus on the discussion of approaches tackling a classification into a larger number of concepts such as [3], [20], [23] and [27].

Hahn and Schnattinger [27] create a *hypothesis space* when encountering an unknown word in a text for each concept that the word could belong to. These initial hypothesis spaces are then iteratively refined on the basis of evidence extracted from the linguistic context the unknown word appears in. In their approach, evidence is formalized in the form of quality labels attached to each hypothesis space. At the end the hypothesis space with maximal evidence with regard to the qualification calculus used is chosen as the correct ontological concept for the word in question. The results of the different version of Hahn et al's system (compare [27]) in terms of accuracy can be found in Table 2. Their approach is very related to ours and in fact they use similar patterns to identify instances from the text. However, the approaches cannot be directly compared. On the one hand they tackle categorization into an even larger number of concepts than we do and hence our task would be easier. On the other hand they evaluate their approach under clean room conditions as they assume accurately identified syntactic and semantic relationships and an elaborate ontology structure, while our evaluation is based on very noisy real-world input — rendering our task harder than theirs.

Alfonseca and Manandhar [3] have also addressed the problem of assigning the correct ontological class to unknown words. Their system is based on the distributional hypothesis, i.e. that words are similar to the extent to which they share linguistic contexts. In this line, they adopt a vector-space model and exploit certain syntactic dependencies as features of the vector representing a certain word. The unknown word is then assigned to the category corresponding to the most similar vector. The best result measured against a reference standard (strict evaluation mode as they call it) was achieved using only verb/object dependencies as features (compare Table 2). Their results seem thus lower compared to our system, but they are also considering a much larger number of concepts, i.e. 1200.

Fleischmann and Hovy [23] address the classification of named entities into fine-grained categories. In particular, they categorize named entities denoting persons into the following 8 categories: *athlete*, *politician/government*, *clergy*, *busi-*

System	No. Concepts	Preprocessing	Accuracy/Recall
MUC	3	various	>90%
Fleischman et al.	8	N-gram frequency extraction	70.4%
Evans	2-8	typology derivation (clustering)	41.41%
PANKOW	59	none	24.9%
Hahn et al. (Baseline)	325	perfect syntactic and semantic analysis	21%
Hahn et al. (TH)	325	perfect syntactic and semantic analysis	26%
Hahn et al. (CB)	325	perfect syntactic and semantic analysis	31%
Alfonseca et al. (Object)	1200	syntactic analysis	17.39%

Table 2: Comparison of results

nessperson, *entertainer/ artist*, *lawyer*, *doctor/scientist*, *police*. Given this categorization task, they present an experiment in which they examine 5 different Machine Learning algorithms: C4.5, a feed-forward neural network, k-nearest Neighbors, a Support Vector Machine and a Naive Bayes classifier. As features for the classifiers they make use of the frequencies of certain N-grams preceding and following the instance in question as well as topic signature features which are complemented with synonymy and hypernym information from WordNet. They report a best result of an accuracy of 70.4% when using the C4.5 decision tree classifer. Fleischman and Hovy's results are certainly very high in comparison to ours – and also to the ones of Hahn et al. [27] and Alfonseca et al. [3] – but on the other hand though they address a harder task than the MUC Named Entity Task, they are still quite away from the number of categories we consider here.

Evans [20] derives similar statistical fingerprints as considered in our approach by querying Google^{TM} and then clusters named entities on the basis of these fingerprints as features in order to derive a class topology from the document in question. He uses a bottom-up hierarchical clustering algorithm for this purpose. His approach differs from the others discussed here in that it is totally unsupervised without even the set of categories being given. Thus, the entities are classified with respect to different sets of categories depending on the document considered. Overall, he reports 41.41% of correctly classified entities, considering from 2 to 8 classes.

In the field of ontology learning, researchers have been using on the one hand unsupervised *context-based* approaches. Maedche et al. [34] for example use a k-nearest neighbours approach to classify an unknown concept into an existing ontology. Caraballo [10], Faure et al. [21] as well as Bisson et al. [6] use bottom-up hierarchical clustering techniques to learn concept hierarchies. Cimiano et al. [14] present an approach based on Formal Concept Analysis and compare it to hierarchical agglomerative clustering and Bi-Section-KMeans as an instance of a partitional algorithm. The problem of these approaches seems certainly that the quality of the automatically acquired ontologies seems low.

On the other hand, there is quite a lot of work related to the use of linguistic patterns to discover certain ontological relations from text. Hearst's [29] seminal work had the aim of discovering taxonomic relations from electronic dictionaries. The precision of the *isa*-relations learned is 61/106 (57.55%) when measured against WordNet as gold standard. Hearst's idea has been reapplied by different researchers with either slight variations in the patterns used [31], in very specific domains [2], to acquire knowledge for anaphora resolution [37], or to discover other kinds of semantic relations such as part-of relations [12] or causation relations [25].

Instead of matching these patterns in a large text collection, some researchers have recently turned to the Web to match these patterns such as in [13], [17], [36]. Some researchers have also used the World Wide Web for question answering purposes such as in [4], [33] or [39], for discovering synonyms [44] or to avoid data sparseness problems [1; 26; 32].

Especially interesting in our context is the work in [17], which aim is to acquire instances for a given concept. In particular, Etzioni et al. present results on the task of acquiring instances of cities, countries, US states, films and actors. In contrast to our approach, they actually download the pages and match the patterns locally instead of generating patterns and counting their hits, thus creating less network traffic than with our approach. Interestingly, they also make use of a Bayesian classifier in order to decide weather an instance belongs to a certain concept or not. Recently, they have also considered learning new patterns by a rule induction process [19]. Though our approaches are definitely related, the aims are to some extent orthogonal. While we aim at classifying a given concept or instance, Etzioni et al. aim at learning the extension of certain concepts for use within a search engine which 'knows it all'.

Brin [8] presents a bootstrapping approach in which the system starts with a few patterns, and then tries to induce new patterns using the results of the application of the seed patterns as training dataset. This is also the general idea underlying the Armadillo system [16], which exploits redundancy in the World Wide Web to induce such extraction rules.

Before concluding this section on related work it seems important to mention that any approach exploiting the Web to discover redundancies or overcome data sparseness faces inherent limits. Brewster et al. [7] for example have argued that in the Web a lot of information remains implicit in the head of web page creators, forming part of their background knowledge and never expressed in an explicit way. This inherent problem of a non-technical nature seems difficult to overcome and gets certainly more important the more technical the domain of consideration becomes.

5. CONCLUSION

We have proposed a new methodology to overcome the *knowledge acquisition bottleneck*. The core of this methodology is a two-stage process in which first collective knowledge about certain items is collected and then presented to a knowledge engineer to be applied in a specific application context. We have also presented a concrete instantiation of this methodology, PANKOW, in which collective knowledge is acquired by matching specific lexico-syntactic patterns in the World Wide Web, leading to the creation of so called statistical fingerprints. These are presented to the knowledge engineer in the form of certain snapshots or views to support him in the creation of metadata and knowledge. Further, we have presented an evaluation of PANKOW with respect to two tasks, one consisting in the classification of instances with regard to an existing ontology and one with the aim of finding the appropriate superconcept for a given concept. In both tasks the results are very promising, especially the ones for the interaction mode in which the knowledge engineer gets presented the top-5 best predictions of the system, which clearly corroborates the practical usefulness of our two-stage methodology.

In our methodology, collective knowledge may be however dominated by a context different from the one in which the given entity to be classified appears. By always classifying entities with respect to the concept with maximal number of hits in the statistical fingerprint, we are thus actually creating a bias towards senses which are predominant in the Web. In future work we will address this issue by attempting to provide more context-based classifications by taking into account the similarity between the page to be annotated and the page in which the pattern was matched, thus hopefully increasing the accuracy of our approach. Moreover, ambiguity is handled only implicitly through the fact that the statistical fingerprint contains all the concepts the entity could possibly be classified with. However, a more explicit and systematic treatment of ambiguity also taking into account the fact that the relation between words and concepts is not one-to-one is certainly desirable. Further, we will also tackle the issue of scalability. Finally, instead of issuing such a large amount of queries to theGoogle^{TM} API, we will examine the possibility of actually downloading the abstracts of the pages and processing them offline, thus considerably reducing network traffic.

Acknowledgements

This work started as a cooperation with Siegfried Handschuh; we acknowledge his support, feedback and help with the initial experiments and the evaluation of the system. We would also like to thank Google^{TM} for support with their Web service API. Thanks to Kai Kühn for help with implementing Pankow and integrating it with the *OntoMat Annotizer* annotation tool. Further, we would also like to acknowledge Ursula Sciesinski and Laura Goebbes for annotating the Lonely Planet pages. Thanks also to our colleagues for their support as well to the reviewers of the WWW 2004 conference for valuable feedback. Research for this paper has been funded by IST-Dot.Kom (grant IST-2001-34038). The Dot.Kom project (http://www.dot-kom.org) is sponsored by the EC as part of the framework V.

6. REFERENCES

[1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the WWW. In *Proceedings of the ECAI Ontology Learning Workshop*, 2000.

[2] K. Ahmad, M. Tariq, B. Vrusias, and C. Handy. Corpus-based thesaurus construction for image retrieval in specialist domains. In *Proceedings of the 25th European Conference on Advances in Information Retrieval (ECIR)*, pages 502–510, 2003.

[3] E. Alfonseca and S. Manandhar. Extending a lexical ontology by a combination of distributional semantics signatures. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002)*, pages 1–7, 2002.

[4] M. Banko, E. Brill, S. Dumais, and J. Lin. AskMSR: Question answering using the Worldwide Web. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, 2002.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

[6] G. Bisson, C. Nedellec, and L. Canamero. Designing clustering methods for ontology building - The Mo'K workbench. In *Proceedings of the ECAI Ontology Learning Workshop*, pages 13–19, 2000.

[7] C. Brewster, F. Ciravegna, and Y. Wilks. Background and foreground knowledge in dynamic ontology construction. In *Proceedings of the SIGIR Semantic Web Workshop*, 2003.

[8] Sergey Brin. Extracting patterns and relations from the World Wide Web. In *Proceedings of the WebDB Workshop at EDBT '98*, pages 172–183, 1998.

[9] M.E. Califf and R.J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *Machine Learning Research*, 4(2):177–210, 2004.

[10] S.A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126, 1999.

[11] J. Carletta. Asessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[12] E. Charniak and M. Berland. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 57–64, 1999.

[13] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proceedings of the 13th World Wide Web Conference*, pages 462–471, 2004.

[14] P. Cimiano, A. Hotho, and S. Staab. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence*, pages 435–439, 2004.

[15] F. Ciravegna. Adaptive information extraction from text by rule induction and generalization. In *Proceedings of tht 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 1251–1256, 2001.

[16] F. Ciravegna, A. Dingli, D. Guthrie, and Y. Wilks. Integrating Information to Bootstrap Information Extraction from Web Sites. In *Proceedings of the IJCAI Workshop on Information Integration on the Web*, pages 9–14, 2003.

[17] H. Cui, M.-Y. Kan, and T.-S. Chua. Unsupervised learning of soft patterns for generating definitions from online news. In *Proceedings of the 13th World Wide Web Conference*, pages 90–99, 2004.

[18] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. Web-scale information extraction in KnowItAll (preliminary results). In *Proceedings of the 13th World Wide Web Conference*, pages 100–109, 2004.

[19] O. Etzioni, M. Cafarella, D. Downey, A-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. Methods for domain-independent information extraction from the web: An experimental comparison. In *Proceedings of the AAAI Conference*, pages 391–398, 2004.

[20] R. Evans. A framework for named entity recognition in the open domain. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP-2003)*, pages 137–144, 2003.

[21] D. Faure and C. Nedellec. A corpus-based conceptual clustering method for verb frames and ontology. In P. Velardi, editor, *Proceedings of the LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, pages 5–12, 1998.

[22] Dieter Fensel. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, 2003.

[23] M. Fleischman and E. Hovy. Fine grained classification of named entities. In *Proceedings of the 19th Conference on Computational Linguistics (COLING)*, 2002.

[24] F. Freitag and N. Kushmerick. Boosted Wrapper Induction. In *Proceedings of AAAI conference*, pages 577–583, 2000.

[25] R. Girju and M. Moldovan. Text mining for causal relations. In *Proceedings of the FLAIRS Conference*, pages 360–364, 2002.

[26] G. Grefenstette. The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB'99 Translating and the Computer 21*, 1999.

[27] U. Hahn and K. Schnattinger. Towards text knowledge engineering. In *AAAI'98/IAAI'98 Proceedings of the 15th National Conference on Artificial Intelligence and the 10th Conference on Innovative Applications of Artificial Intelligence*, pages 524–531, 1998.

[28] S. Handschuh and S. Staab. CREAM - Creating Metadata for the Semantic Web. *Computer Networks*, 42:579–598, 2003.

[29] M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[30] L. Hirschman and N. Chinchor. Muc-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, 1997.

[31] L.M. Iwanska, N. Mata, and K. Kruger. Fully automatic acquisition of taxonomic knowledge from large corpora of texts. In L.M. Iwanksa and S.C. Shapiro, editors, *Natural Language Processing and Knowledge Processing*, pages 335–345. MIT/AAAI Press, 2000.

[32] F. Keller, M. Lapata, and O. Ourioupina. Using the web to overcome data sparseness. In *Proceedings of EMNLP-02*, pages 230–237, 2002.

[33] C. T. Kwok, O. Etzioni, and Daniel S. Weld. Scaling question answering to the web. In *ACM Transactions on Information Systems 2001*, pages 150–161, 2001.

[34] A. Maedche, V. Pekar, and S. Staab. Ontology learning part one - on discovering taxonomic relations from the web. In *Web Intelligence*, pages 301–322. Springer Verlag, 2002.

[35] A. Maedche and S. Staab. Measuring similarity between ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW)*, pages 251–263. Springer Verlag, 2002.

[36] K. Markert, N. Modjeska, and M. Nissim. Using the web for nominal anaphora resolution. In *EACL Workshop on the Computational Treatment of Anaphora*, 2003.

[37] M. Poesio, T. Ishikawa, S. Schulte im Walde, and R. Viera. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, 2002.

[38] M. Poesio and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.

[39] D.R. Radev, H. Qi, Z. Zheng, S. Blair-Goldensohn, Z. Zhang, W. Fan, and J.M. Prager. Mining the web for answers to natural language questions. In *Proceedings of the Conference on Information and Knowledge Management*, pages 143–150, 2001.

[40] P. Resnik and N. Smith. The web as a parallel corpus. *Computational Lingusitics*, 29(3):349–380, 2003.

[41] Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272, 1999.

[42] S. Staab, C. Braun, I. Bruder, A. Düsterhöft, A. Heuer, M. Klettke, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger. Getess - searching the web exploiting german texts. In *Proceedings of the 3rd Workshop on Cooperative Information Agents*, pages 113–124. Springer Verlag, 1999.

[43] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday Books, 2004.

[44] P.D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML)*, pages 491–502, 2001.