Learning Taxonomic Relations from Heterogeneous Sources of Evidence

Philipp Cimiano [a] Aleksander Pivk [b] Lars Schmidt-Thieme [c] Steffen Staab [d]

[a] *Institute AIFB, University of Karlsruhe*
[b] *Jozef Stefan Institute, Ljubljana*
[c] *Computer Based New Media Group, University of Freiburg*
[d] *Institute for Computer Science, University of Koblenz*

Abstract. We present a novel approach to learning taxonomic relations between terms by considering multiple and heterogeneous sources of evidence. In order to derive an optimal combination of these sources, we exploit a machine-learning approach, representing all the sources of evidence as first-order features and training standard classifiers. We consider in particular different features derived from WordNet, an approach matching Hearst-style patterns in a corpus and on the Web as well as further methods mentioned in the literature. In particular, we explore different classifiers as well as various strategies for dealing with unbalanced datasets. We evaluate our approach by comparing the results with a reference taxonomy for the tourism domain.

Keywords. ontology learning, taxonomies, text and web mining, machine learning

1. Introduction

Taxonomies, thesauri or concept hierarchies are a crucial component of many applications within the Semantic Web [2], Knowledge Management [11], Information Retrieval [1,26], Text Clustering [17], Natural Language Processing and Information Systems in general. In fact, there has been a long tradition in Artificial Intelligence and related fields such as Natural Language Processing or Information Retrieval to automatically learn taxonomies from data. As text documents are massively available, most researchers have attempted to learn taxonomies on the basis of textual input, whereby mainly three different learning paradigms have been exploited. First, several researchers have attempted to find taxonomic relations expressed more or less explicitly in texts by matching certain patterns which we will refer to as Hearst-patterns in line with Hearst's seminal work [15] and follow-up work in [1], [6], [18] and [22]. Some researchers have even went further and searched for these patterns on the Web [7,9,20]. Other researchers have used the internal structure of noun phrases to find taxonomic relations [4,29]. Second, many researchers follow Harris' distributional hypothesis basically claiming that words or terms are semantically similar to the extent to which they share similar syntactic contexts [14]. The most prominent examples of this approach are probably [3], [5], [10], [12], [16], [21], and more recently also [8] and [24]. Third and finally, there are also approaches relying on a document-based notion of term subsumption such as the one found in [26].

However, there has been certainly almost no work on combining different learning paradigms. The aim of this paper is to present an approach which combines the three above mentioned paradigms as well as a few other approaches and resources to learn taxonomic relations from all these heterogeneous sources. The crucial question herein is thus to find an optimal combination of the indications provided by all these approaches. As any manual attempt to combine these different approaches would certainly be *adhoc*, we resort to a supervised scenario in which an optimal combination is learned from the data itself and make use of standard classifiers for this purpose. In fact, we learn classifiers which given two terms as well as the results of all the different approaches considered, decide if they stand in a taxonomic relation or not. As most of the terms in a given taxonomy do not stand in such a relation, we are thus faced with very unbalanced datasets making it necessary to apply strategies to cope with such skewed distributions as described in [23].

In this paper we thus examine the possibility of learning taxonomic relations by combining the evidence from different sources and techniques using a classification approach. The crucial questions we address in this paper are (i) how to convert the different sources of evidence and results of different approaches into first-order features which can be used by a classifier, (ii) which classifiers perform best on the task and (iii) which strategies are most suitable to deal with the unbalanced datasets we consider.

The paper is structured as follows: in Section 2 we describe our dataset, i.e. the corpus we use, the ontology we aim at reproducing as well as discuss our evaluation strategy. In Section 3 we present the different sources of evidence we consider and in Section 4 we present some results. Before concluding, we discuss some related work in Section 5.

2. Dataset and Evaluation

As underlying corpus for the corpus-based sources of information we use two domain-specific text collections: a collection of texts from *http://www.lonelyplanet.com* as well as from *http://www.all-in-all.de*, a site containing information about accommodation, activities etc. of *Mecklenburg Vorpommen*, a region in northeast Germany. Furthermore, we also use a general corpus, the British National Corpus. Altogether the corpus size is over 118 Million tokens.

The concept hierarchy or taxonomy we consider for evaluating our approach is a tourism reference ontology modeled by an experienced ontology engineer within the GETESS project [28]. The ontology is rather small with 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 non-direct is-a relations between them. For our evaluation we take into account the set of direct and non-direct *isa* relations. In particular, we evaluate the is-a relations found by our system with the direct and non-direct ones in terms of Recall, Precision and F-Measure. It is important also to mention that we consider only pairs of terms/concepts contained in the concept hierarchy, which we thus aim at ’reproducing’ with our approach.

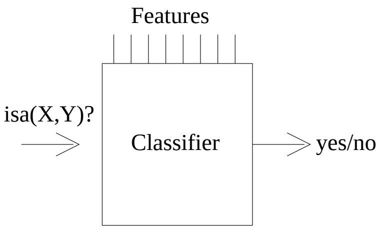

Figure 1. Learning from Heterogeneous Sources of Evidence

3. Learning From Heterogeneous Evidence

In this section we describe the different sources of evidence we aim at combining via our classification-based approach to taxonomy learning. In particular, we discuss how the various sources can be transformed into first-order features which can be used by a classifier.

3.1. Hearst Patterns

3.1.1. Matching Patterns in a Corpus

The first source of evidence we consider are lexico-syntactic patterns matched in a certain corpus in line with [15], where the patterns we use are mainly taken from:[1]

(1) NP_0 such as NP_1, NP_2, ..., NP_{n-1} (and|or) NP_n[2]
(2) such NP_0 as NP_1, NP_2, ... NP_{n-1} (and|or) NP_n
(3) NP_1, NP_2, ..., NP_n (and|or) other NP_0
(4) NP_0, (including|especially) NP_1, NP_2, ..., NP_{n-1} (and|or) NP_n
(5) NP_1 is NP_0
(6) NP_1, another NP_0
(7) NP_0 like NP_1

According to Hearst, from such patterns we can derive that for all $NP_i,\ 1 \leq i \leq n,\ is\text{-}a(head(NP_i), head(NP_0))$[3]. Now given two terms t_1 and t_2 we record how many times a Hearst-pattern indicating an *isa*-relation between t_1 and t_2 is matched in the corpus. We then normalize this value by dividing by the maximum number of Hearst patterns found for t_1. In order to match the above patterns we create regular expressions over part-of-speech tags to match NP's. In particular, we use the tagger described in [27] and match non-recursive NP's consisting of a determiner, an optional sequence of modifying adjectives and a sequence of common nouns constituting the head of the NP. For the *conference* concept for example, we find the following candidate *isa* relations, where the number in the second column gives the normalized value as described above:

[1] Pattern 7 is taken from [18].

[2] NP_i stands for a noun phrase.

[3] Actually [15] states that for all $NP_i,\ 1 \leq i \leq n$, $hypernym(head(NP_i), head(NP_0))$, but we raise terms to the status of concepts – thus neglecting polysemy – and go one step further stating that a Hearst pattern is an indicator for an *is-a*-relationship which from a formal point of view is interpreted as *subsumption* in most ontology formalisms.

isa$_{Hearst}$(conference,event)	0.36
isa$_{Hearst}$(conference,body)	0.18
isa$_{Hearst}$(conference,course)	0.09
isa$_{Hearst}$(conference,weekend)	0.09
isa$_{Hearst}$(conference,meeting)	0.09
isa$_{Hearst}$(conference,activity)	0.09

The first interesting observation here is that, despite of using quite a big corpus, Hearst patterns appear relatively rarely. When using only Hearst's approach and classifying a pair of terms as *isa*-related if isa_{Hearst} is above a threshold t, we get the best F-Measure of $F = 10.64\%$ at $t = 0.03$ with a precision of $P = 25\%$ and a recall of $R = 6.76\%$. The best precision was 60% using a threshold of $t = 0.5$.

3.1.2. Generating Patterns

Certainly, when using a corpus we have to cope with typical data sparseness problems. On the other hand, some researchers have shown that the World Wide Web is an attractive way of reducing data sparseness [13,19,25]. In this line, following [20], we use the Google API[4] to count the matches of a certain expression in the Web. In particular, for each pair $(t_1$,$t_2)$, we generate the following patterns and count the number of hits returned by the Google API:

$\pi(t_1)$ such as $\pi(t_2)$
such $\pi(t_1)$ as $\pi(t_2)$
$\pi(t_1)$, including $\pi(t_2)$
$\pi(t_1)$, especially $\pi(t_2)$
$\pi(t_2)$ and other $\pi(t_1)$
$\pi(t_2)$ or other $\pi(t_1)$

where $\pi(t)$ returns the correct plural form of t.
As above, these patterns are indicators for a corresponding taxonomic relation isa$_{WWW}(t_1$,$t_2)$. So, this source of evidence is certainly similar in spirit to the Hearst approach described above, but with the main difference that above the patterns are matched against a corpus and here for each pair $(t_1$,$t_2)$ a certain number of patterns are generated and then sent as queries to the Google API. The sum of the number of Google hits over all patterns for a certain pair (t_1, t_2) is then normalized by dividing through the number of hits returned for t_1. Here are the top five matches for the *conference* concept and other terms in the tourism concept hierarchy we consider; the value in the second column indicates the normalized number of hits returned by the Google API:

isa$_{WWW}$(conference,service)	0.27
isa$_{WWW}$(conference,event)	0.25
isa$_{WWW}$(conference,area)	0.11
isa$_{WWW}$(conference,organization)	0.05
isa$_{WWW}$(conference,information)	0.04

It is important to note that due to the simple morphology we have used, we get no information for nouns which do not form their plural regularly, e.g. *activity*. The best F-

[4] http://www.google.com/apis/

Measure here was F=18.84% with a precision of P=15.77% and a recall of R=23.43% when selecting all the relations above a threshold of 0.04. So here we yield a greater recall at the cost of also a lower precision which is due to the fact that the WWW is a very general resource and the pattern-matching approach also yields a considerable amount of errors.

3.1.3. *Downloading Web Pages*

Furthermore, as an alternative to the pattern generation approach described above, we also follow an approach in which web pages are actually downloaded and Hearst patterns are matched offline thus overcoming the idiosyncrasies with the generation of plural forms and also allowing to match expressions with a more complex linguistic structure. For this purpose, we assign one or more functions $f_i : string \rightarrow string$ – which we will refer to as clues – to each of the Hearst patterns i to be matched. Given a concept of interest c, we instantiate each of the clues and download a number of pages matching the query $f_i(c)$ using the Google API. For example, given the clue $f(x) = ”such\ as” \oplus \pi(x)$ and the concept *conference* we would download 100 abstracts matching the query f(conference), i.e. ”such as conferences”.[5] For each concept of interest and for each of the correspondingly instantiated clues, we then process the downloaded abstracts by matching the corresponding pattern, thus yielding its potential superconcepts. As described above, for each pair t_1, t_2 we calculate the number of times t_1 and t_2 were found to stand in an *isa*-relation divided by the number of times t_1 was matched in a pattern as subconcept, i.e. $isa_{WWW'} = \frac{|patterns(t_1,t_2)|}{|patterns(t_1,*)|}$. The following table gives the clues used as well as the corresponding Hearst patterns:

Clue	Hearst pattern
$f(x) = ”such\ as” \oplus \pi(x)$	(1)
$f(x) = \pi(x) \oplus ”and\ other”$	(3)
$f(x) = \pi(x) \oplus ”or\ other”$	(3)
$f(x) = ”including” \oplus \pi(x)$	(4)
$f(x) = ”especially” \oplus \pi(x)$	(4)
$f(x) = x \oplus ”is”$	(5)

The top four pairs for the *conference* concept were:

isa$_{WWW'}$(conference,event)	0.27
isa$_{WWW'}$(conference,activity)	0.17
isa$_{WWW'}$(conference,initiative)	0.03
isa$_{WWW'}$(conference,function)	0.03

Here using the simple threshold-classifier we get an F-Measure of $F = 17.58\%$ with a precision of $P = 16.12\%$ and a recall of $R = 19.34\%$.

3.2. *WordNet*

As a further source of evidence we use the hypernymy information from WordNet[6]. Actually, WordNet can not be seen as an unstructured source of evidence, but the informa-

[5] Here, $\oplus$ denotes the concatenation operator defined on two strings.

[6] We used version 1.7.1 for our experiments.

tion contained in it is so general and domain independent that when exploiting it in the context of a specific domain, it has to be treated as an uncertain source of evidence such as the other sources we consider here. So, given two terms t_1 and t_2, we check if they stand in a hypernym relation with regard to WordNet. It is important to note that two terms t_1 and t_2 can appear in more than one synset and thus there could be more than just one 'hypernymic' path from the synsets of t_1 to the synsets of t_2. Here we normalize the number of hypernymic paths by dividing by the number of senses of t_1, setting 1 as maximum, i.e. we consider the value $isa_{WN}(t_1, t_2) = max(\frac{|paths(senses(t_1),senses(t_2))|}{|senses(t_1)|}, 1)$. For example, in WordNet there are four such different 'hypernymic' paths between the synsets of *country* and the ones of *region*. Furthermore, *country* has 5 senses and so this value would be 0.8. For *conference*, which has 3 senses in WordNet, we get the following candidate taxonomic relations:

isa$_{WN}$(conference, organization)	1
isa$_{WN}$(conference,group)	0.67

Further, we also consider a variant of taking into account the WordNet hierarchy in which we consider only the first, i.e. most frequent, sense of t_1 as specified by the formula $isa_{WN_{first}} = max(|paths(first\_sense(t_1), senses(t_2))|, 1)$. This value is obviously 0 or 1. The precision for the *isa* pairs extracted from WordNet is much lower than for the ones from the Hearst patterns which is due to the fact that WordNet contains so much ambiguity and it is domain independent. The precision is in fact around P=21.6% when considering all senses and regarding all relations with a value above 0.2 as correct and around P=30.55% when taking into account only the first sense. While the recall is higher than with Hearst's approach, it is still quite low at R=7.23% and R=5.19%, respectively. The best F-Measure for the feature considering all senses is thus F=10.84% and F=8.87% for the feature considering only the first sense.

It is important to emphasize that this does not mean that the relations found in WordNet are totally wrong, but that they do not appear in our target ontology. After manual inspection of the relations in WordNet and the ones in the target ontology we found that certain terms are modeled in a very different manner, which explains why the precision of the relations found in WordNet is so low when compared with the target hierarchy. For example, according to WordNet, *presentation* is a *human activity* (most frequent sense), while according to our target ontology, *presentation* is a *business event*. Another example here is *night*, which according to WordNet is a *period* and according to our target ontology is a *time*. Further, according to WordNet, *price list* is an *information*, while according to our target ontology *price list* is an *agreement*.

3.3. 'Head Matching'-Heuristic

In order to identify further *isa* relations, we make use of a heuristic used by [29] which we will henceforth call *'head matching'*-heuristic. Basically, given two terms t_1 and t_2, if t_2 matches t_1 and t_1 is additionally modified by certain terms or adjectives, we derive the relation isa(t_1,t_2). As an example, according to this heuristic, we might derive that t_1='international conference' and t_2='conference' are related by an *isa* relation, i.e. isa$_{head}$(international conference,conference). This is similar to the *HeadNounToClass_ModToSubClass* rule described in [4]. When evaluating this heuristic on our dataset, we get a precision of 50%, a very low recall of 3.77% and an F-Measure of F=7.02%.

3.4. Corpus-based subsumption

As a further source of evidence we also introduce a corpus-based notion of subsumption and regard a term t_1 as a subclass of t_2 if all the syntactic contexts in which t_1 appears are also shared by t_2. For this purpose, for each term in question we extract pseudo-syntactic dependencies from the corpus. These dependencies are not really syntactical as they are not obtained from parse trees, but with a very shallow method consisting in matching certain regular expressions over part of speech tags. The motivation for doing this is the observation in [12] that the quality of using word windows or syntactic dependencies for distributional analyses depends on the rank or frequency of the word in question. In this line, our intention is to make a compromise between using word windows and syntactic dependencies extracted from parse trees. Our pseudo-syntactic dependencies are surface dependencies extracted by matching regular expressions. In what follows we list the syntactic expressions we use and give a brief example of how the features are extracted from these expressions:

- adjective modifiers, i.e. *a nice city* → nice ∈ features(city)
- prepositional phrase modifiers, i.e. *a city near the river* → near_river ∈ features(city) and city_near ∈ features(river), respectively
- possessive modifiers, i.e. *the city's center* → has_center ∈ features(city)
- noun phrases in subject or object position. i.e. *the city offers an exciting nightlife* → offer_subj ∈ features(city) and offer_obj ∈ features(nightlife)
- prepositional phrases following a verb, i.e. *the river flows through the city* → flows_through ∈ features(city) and flows_subj ∈ features(river)
- copula constructs i.e. *a flamingo is a bird* → is_bird ∈ features(flamingo)
- verb phrases with the verb *to have*, i.e. *every country has a capital* → has_capital ∈ features(country)

Consider for example the following discourse:

Mopti is the biggest city along the Niger with one of the most vibrant ports and a large bustling market. Mopti has a traditional ambience that other towns seem to have lost. It is also the center of the local tourist industry and suffers from hard-sell overload. The nearby junction towns of Gao and San offer nice views over the Niger's delta.

Here we would extract the following terms and features:

Term	Features
city	biggest
ambience	traditional
center	of_tourist_industry
junction town	nearby
market	bustling
port	vibrant
tourist industry	center_of, local
overload	suffer_from
town	seem_subj
view	nice, offer_obj

On the basis of these term vectors we calculate a directed Jaccard coefficient as follows: $isa_{corpus}(t_1, t_2) = \frac{|features(t_1) \cap features(t_2)|}{|features(t_1)|}$, thus computing the number of common features divided by the number of features of term t_1. So, the measure presented here gives a normalized value between [0..1] indicating in how far $features(t_1)$ is included in $features(t_2)$.
Here follow the top ten superconcepts for *conference* according to this method:

isa$_{corpus}$(conference,congress)	0.44
isa$_{corpus}$(conference,seminar)	0.44
isa$_{corpus}$(conference,masseur)	0.43
isa$_{corpus}$(conference,banquet)	0.34
isa$_{corpus}$(conference,aerobic)	0.37
isa$_{corpus}$(conference,pilgrimage)	0.33
isa$_{corpus}$(conference,elevator)	0.31
isa$_{corpus}$(conference,sanatorium)	0.31
isa$_{corpus}$(conference,brochure)	0.30
isa$_{corpus}$(conference,cabaret)	0.30

Evaluated on our reference taxonomy, the simple threshold classifier yielded a relatively high recall of $R = 27.83\%$ but a very low precision and F-Measure of $P = 0.92\%$ and $F = 1.78\%$, respectively.

3.5. *Document-based Subsumption*

Sanderson and Croft [26] have suggested a document-based notion of subsumption according to which a term t_1 is a subclass of term t_2 if t_2 appears in all documents in which t_1 appears. Instead of computing these results with respect to a corpus we resort once more the the World Wide Web and use the Google API to calculate the number of documents in which t_1 and t_2 occur, dividing this value by the number of documents in which t_1 occurs. Thus we also yield a value between [0..1]. According to these document coocurrence method, the top ten superconcepts for *conference* are:

isa$_{croft}$(conference,information)	0.17
isa$_{croft}$(conference,service)	0.17
isa$_{croft}$(conference,day)	0.16
isa$_{croft}$(conference,time)	0.16
isa$_{croft}$(conference,email)	0.15
isa$_{croft}$(conference,event)	0.14
isa$_{croft}$(conference,date)	0.14
isa$_{croft}$(conference,area)	0.12
isa$_{croft}$(conference,place)	0.12
isa$_{croft}$(conference,organization)	0.11

Here the best result of the threshold classifier yielded an F-Measure of $F = 6.32\%$ at a precision of $P = 13.98\%$ and a recall of $R = 4.09\%$.

4. Results

As classifiers we use a Naive Bayes (NB) classifier, a C4.5 decision tree classifier, a Perceptron (PER) as well as a Multi-layer Perceptron (MLPER) with one hidden layer consisting of as many hidden nodes as input nodes. We use the version of these algorithms implemented in WEKA[7] using standard settings and give results averaged over ten runs. In particular, we use 60% of the dataset for training and 40% for testing. Further, in order to address the problem of the unbalanced dataset, we experiment with the following strategies: (i) undersampling [23], (ii) oversampling [23], (iii) varying the classification threshold as well as (iv) introducing a cost matrix. Additionally, we also report on results of experimenting with one-class Support Vector Machines, for which we obviously need not to worry about the unbalanced character of the dataset as they merely make use of positive examples for training.

4.1. Baselines

As already mentioned above, in order to evaluate our machine learning approach, for each feature we calculated the results with respect to our dataset of a very simple classifier assigning an example to the *isa* class if the value of the corresponding feature is above a threshold t. For each feature we varied the threshold from 0 to 1 in steps of 0.01. The F-Measure, precision and recall values for the best threshold parameter t for each feature are summarized in the table below. Further, as a very simple combination of the features we experimented with two further classifiers assigning an example to the *isa* class if the average or the maximum of the values of features 1-6[8] is above a threshold t (compare the results in the table below).

No.	Feature	t	F	P	R
1	isa_{Hearst}	0.03	10.64%	25%	6.76%
2	isa_{WWW}	0.04	18.84%	15.77%	23.43%
3	$isa_{WWW'}$	0	17.58%	16.12%	19.34%
4	isa_{WN}	0.2	10.84%	21.60%	7.23%
5	$isa_{WN_{first}}$	0	8.87%	30.55%	5.19%
6	$isa_{vertical}$	0	7.02%	50%	3.77%
7	isa_{corpus}	0.01	1.78%	0.92%	27.83%
8	isa_{croft}	0.6	6.32%	13.98%	4.09%
	Average(1-6)	0.02	21.28%	18.61%	24.84%
	Maximum(1-6)	0.12	21%	19.03%	23.43%

4.2. Undersampling

Undersampling (compare [23]) consists in removing a number of examples of the majority class, in our case the *non-isa* examples or, which is equivalent, to select only a subset of the examples of the majority class for training. In our experiments we randomly selected a number of negative examples which equals the number of positive examples multiplied by an undersampling factor f_U, i.e. NumberNegatives = f_U * NumberPosi-

[7] http://www.cs.waikato.ac.nz/∼ml/weka/

[8] When adding the features 7 and 8 the results are actually worse.

tives. We varied the factor f_U from 1 to 30. The results for all classifiers are given in Figure 2 which shows the F-Measure over the undersampling factor f_U. The best F-Measure of $F = 21.50\%$ was obtained with $f_U = 13$ using the Mulitlayer Perceptron (MLPER), thus being slightly over the baseline.

4.3. *Oversampling*

In contrast to undersampling, oversampling consists in adding additional examples of the minority class [23], in our case the *isa* class. In our experiments we randomly selected a number of positive examples equal to the original number of positive examples multiplied with a factor f_O, i.e. NumberPositives = NumberPositivesOriginal * f_O. We varied the oversampling parameter from 0 to 20 in steps of 1. The corresponding results are depicted in Figure 3.[9] With this oversampling strategy we get better results than with the undersampling strategy, achieving an F-Measure of 22.86% using the Multilayer Perceptron (MLPER) and an oversampling factor $f_O = 11$.

4.4. *Threshold*

Another possibility is to vary the classification threshold of the classifier. Almost all classifiers internally compute for each example a probability of belonging to each target class, assigning the example to the class with the highest probability. In our binary case, an example is thus classified as *isa* if this probability is greater than 0.5. We varied also this threshold from 0 to 1 in steps of 0.05. The corresponding results for all the classifiers are depicted in Figure 4. The best F-Measure of $F = 18.7\%$ was achieved using the Multilayer Perceptron (MLPER) and a threshold of 0.1. With this strategy we did thus not improve uppon the baseline.

4.5. *Cost Matrix*

In WEKA it is possible to specify a cost matrix indicating the relative cost of misclassifying an example. In further experiments we made use of this possibility, introducing a factor f_C specifying the relative cost of misclassifying an *isa* example as *non-isa* with respect to misclassifying a *non-isa* as *isa*. We varied this factor from 1 to 10 in steps of 1. The results in terms of F-Measure over this factor are given in Figure 5. Here the best F-Measure of F=20.09% was achieved when using the Multilayer Perceptron and a relative misclassification cost of 6:1. Thus also when using this strategy we do not improve above the baseline.

4.6. *One Class SVM*

Further, we also experimented with one-class Support Vector Machines which do rely only on positive examples for training. Thus, the unbalanced character of the dataset is not an issue here. In particular, we used the regression SVM implementation of the TextGarden tool suite[10] and used standard settings, performing evaluation with n-fold

[9] Unfortunately, for the oversampling as well as for the varying cost strategies (see below), we have not been able to perform our experiments with C4.5 decision trees as WEKA reported not to have enough memory.

[10] http://kt.ijs.si/dunja/TextGarden/

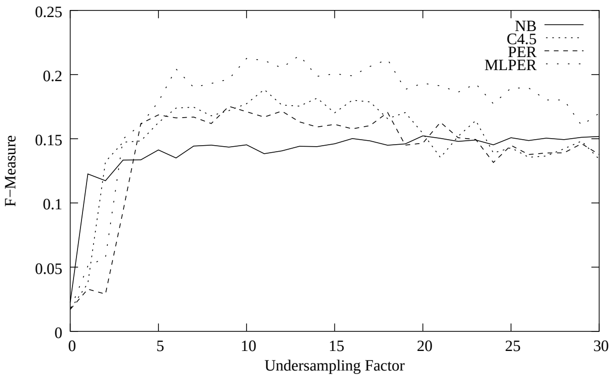

Figure 2. Results for undersampling

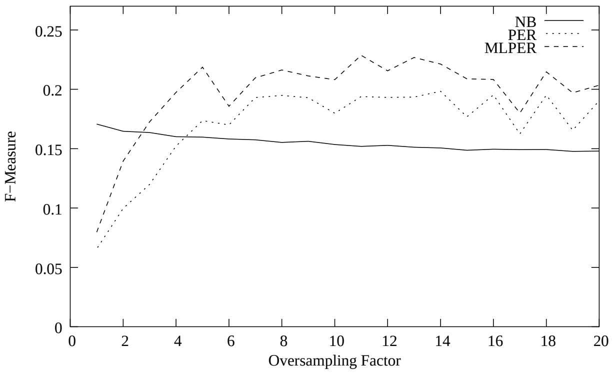

Figure 3. Results for oversampling

Train/Test split	F-Measure	Precision	Recall
1/1	32.72%	36.98%	R=29.38%
2/1	32.96%	37.85%	R=29.21%
3/1	32.38%	37.65%	R=28.47%
4/1	32.91%	37.64%	R=29.35%

Table 1. Results of one-class SVM for different train/test splits

cross validation, where n is the number of data splits. We experimented here with different test/training splits obtaining the best result of $F = 32.96\%$ with a split of 2:1. Table 1 shows the F-Measure, Precision and Recall values for the different splits used.

4.7. Discussion

The best results achieved with the one-class SVM ($F = 32.96\%$) are more than 10 points above the baseline classifier taking into account the average ($F = 21.28\%$) or the maximum ($F = 21\%$) of the different approaches considered. Furthermore, the best result is also more than 14 points better than the best single-feature classifier using the isa_{WWW} feature ($F = 18.84\%$). The results thus show that our supervised approach to combining

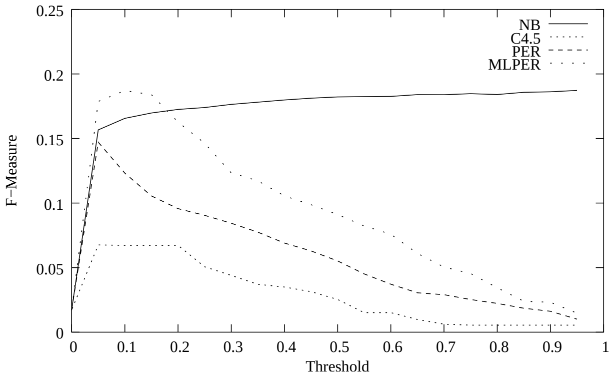

Figure 4. Results for varying the threshold

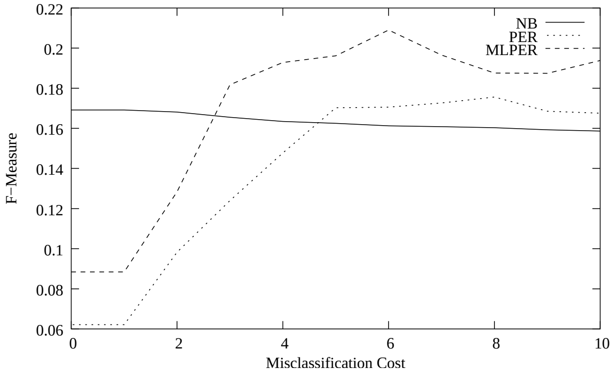

Figure 5. Results for varying the misclassification cost

different indicators from multiple and heterogeneous sources indeed yields very promising results. The second best results were achieved using a Multilayer Perceptron as well as oversampling and undersampling as strategies to cope with the unbalanced character of the dataset. Varying the threshold or the misclassification cost did not yield better results compared to the baseline. Interesting is also a detailed analysis of the weights assigned to the different features by the one-class SVM classifier as it will allow more insight in which features are good predictors of an isa-relation and which ones are not. Figure 6 shows the weight for the different dataset splits used with the SVM classifier. The most reliable predictor of an isa-relation is clearly the 'Head Matching'-heuristic. The second most reliable feature is the approach matching Hearst patterns in the corpus. The third best feature is the version of the WordNet approach using only the first sense. The version of WordNet using all senses as well as both approaches matching Hearst patterns in the Web do not perform as good, but still reasonably. The feature corresponding to the corpus-based subsumption seems to be a good negative indicator, which probably indicates that we should normalize by dividing by the features of t_2 instead of t_1 (see Section 3). The document-based subsumption feature seems to behave neutrally.

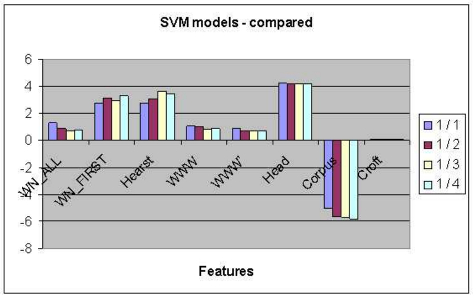

Figure 6. Weight of each feature for the different train/test splits

5. Related Work

Several paradigms have been applied to the problem of learning concept hierarchies from textual data. Many researchers follow Hearst's seminal work [15] who makes use of lexico-syntactic patterns manually acquired from a corpus to identify hyponym/hypernym relations. The approaches of Hearst and others are characterized by a (relatively) high precision in the sense that the quality of the learned relations is high. However, these approaches suffer from a very low recall which is due to the fact that the patterns are very rare in corpora. For this reason, in addition to matching Hearst-style patterns in a corpus, we also exploit the WWW to find more instances of these patterns. Other researchers have made use of the internal structure of noun phrases to derive sub-/superconcept relations such as in [4] or [29]. In our approach we have also included a feature regarding a term t_1 as a subconcept of a term t_2 if t_1 matches t_2 and t_2 is additionally preceded by some modifiers.

The approaches in [3], [5], [8], [10], [16] and [21] make use of clustering techniques relying on the assumption that terms are similar to the extent to which they share syntactic contexts. The directed Jaccard coefficient used in our combination approach is in fact a directed similarity measure relying on contextual overlap of terms.

Interesting is also the document-based notion of subsumption used in [26], where a hierarchy between nouns is derived automatically by considering the document a certain term appears in as context. In particular, Sanderson and Croft present a document-based definition of subsumption according to which a certain term t_1 is more special than a term t_2 if t_2 also appears in all the documents in which t_1 appears. In the approach presented in this paper we have reused this idea by calculating a normalized value indicating how many of the documents containing t_1 also contain t_2 and using this value in our classification-based approach.

None of the above approaches however considers the possibility of learning taxonomic relations by combining different learning paradigms. In this respect our approach is novel and unique, combining in essence all the above mentioned paradigms: (i) the matching of lexico-syntactic patterns indicating a certain semantic relation, (ii) analyzing the internal structure of noun phrases, (iii) the corpus-based assessment of similarity and (iv) the document-based notion of subsumption described above.

6. Conclusion and Outlook

We have presented a novel and original approach to learning taxonomic relations by considering various and heterogeneous sources such as a text corpus, the Web, WordNet, etc. The approach combines state-of-the-art ontology learning paradigms to extract first-order features from the above sources and uses supervised machine-learning techniques to derive an optimal combination of the different features with respect to the task of deciding if two given terms stand in a taxonomic relation or not. We have shown that a successful combination of different sources and techniques indeed improves the results on the task with respect to a simple combination strategy as well as with respect to all the approaches and resources considered on their own.

In general, we see this as a first and important step towards learning complete ontologies. Though the approach presented is supervised, our assumption is that the learned models are to a certain degree domain-independent. Further research should actually clarify whether the approach presented in this paper can be (i) improved by using other classifiers or using additional features, (ii) indeed applied to other domains and (iii) extended to learn other relations than taxonomic ones.

Acknowledgments Philipp Cimiano is currently supported by the IST-Dot.Kom project (http://www.dot-kom.org), sponsored by the EC as part of the framework V, (grant IST-2001-34038). During his stay at the AIFB, Aleksander Pivk has been supported by a Marie Curie Fellowship of the European Community program 'Host Training Sites' and by the Slovenian Ministry of Education, Science and Sport. We would like to acknowledge our colleagues for fruitful discussions. Thanks also to the reviewers of the ECAI Ontology Learning and Population Workshop for valuable comments.

References

[1] 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.

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

[3] 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.

[4] P. Buitelaar, D. Olejnik, and M. Sintek. A protégé plug-in for ontology extraction from text based on linguistic analysis. In *Proceedings of the International Semantic Web Conference (ISWC)*, 2003.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] G. Grefenstette. *Explorations in Authomatic Thesaurus Construction*. Kluwer, 1994.

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

[14] Z. Harris. *Mathematical Structures of Language*. Wiley, 1968.

[15] 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.

[16] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 268–275, 1990.

[17] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proceedings of the IEEE International Conference on Data Mining*, pages 541–544, 2003.

[18] 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.

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

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

[21] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

[22] 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 (LREC)*, 2002.

[23] Foster Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets*, 2000.

[24] M.-L. Reinberger and P. Spyns. Unsupervised text mining for the learning of dogma-inspired ontologies. In P. Buitelaar, P. Cimiano, and B. Magnini, editors, *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press, 2005. this volume.

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

[26] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Research and Development in Information Retrieval*, pages 206–213. 1999.

[27] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[28] 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.

[29] P. Velardi, R. Navigli, A. Cuchiarelli, and F. Neri. Evaluation of ontolearn, a methodology for automatic population of domain ontologies. In P. Buitelaar, P. Cimiano, and B. Magnini, editors, *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press, 2005. this volume.