Learning Taxonomic Relations from Heterogeneous Evidence

Philipp Cimiano[1], Aleksander Pivk[1,2], Lars Schmidt-Thieme[3], Steffen Staab[1]
[1]**Institute AIFB**
University of Karlsruhe
{cimiano,staab}@aifb.uni-karlsruhe.de
[2]**Jozef Stefan Institute**
Ljubljana
aleksander.pivk@ijs.si
[3]**Computer Based New Media Group**
University of Freiburg
lst@informatik.uni-freiburg.de

Abstract. We present a novel approach to the automatic acquisition of taxonomic relations. The main difference to earlier approaches is that we do not only consider one single source of evidence, i.e. a specific algorithm or approach, but examine the possibility of learning taxonomic relations by considering various and heterogeneous forms of evidence. In particular, we derive these different evidences by using well-known NLP techniques and resources and combine them via two simple strategies. Our approach shows very promising results compared to other results from the literature. The main aim of the work presented in this paper is (i) to gain insight into the behaviour of different approaches to learn taxonomic relations, (ii) to provide a first step towards combining these different approaches, and (iii) to establish a baseline for further research.

1 Introduction

A lot of applications are emerging in the NLP community in which taxonomies or conceptual hierarchies are used as background knowledge to accomplish a certain task. In information retrieval (IR) for example, the availability of an ontology for a given domain allows to replace traditional keyword-based approaches by more sophisticated ontology-based search mechanisms such as the one proposed in [23]. Hotho et al. ([18]) show how bag-of-words based approaches to text clustering can be extended by including information derived from concept hierarchies in the document vectors. In Information Extraction (IE), machine learning based systems can use a concept hierarchy to tag text segments at different levels of detail with regard to the hierarchy as well as to produce more concise extraction rules ([8]). Moreover, named entity recognition/classification (NE(R—C)) can be performed with regard to a concept hierarchy as in [12] instead of with regard to a flat set of (only three) categories as in the MUC named entity task ([16]). Resnik ([28]) shows how concept hierarchies can also be used for the resolution of syntactic and semantic ambiguities.

However, it is well known that the development of suitable ontologies is time-consuming, complex and thus a major bottleneck for such ontology-based applications.

In this paper we present an approach in order to partially overcome this bottleneck by automatically acquiring ontological knowledge considering different sources of evidence. In particular, our approach learns taxonomic (also known as *is-a*) relations by considering information from:

- Hearst-patterns (see [14]) matched in a large text corpus
- Hearst-patterns matched in the World Wide Web in line with [22]
- WordNet ([10])
- the 'vertical relations' heuristic used in [33]

These sources of evidence are normalized in order to be comparable and combined by a very simple and naive approach which however shows that combining diverse and heterogeneous sources of evidence indeed leads to better results. The approach is evaluated with regard to a handcrafted ontology for the tourism domain which was developed in the ontology comparison study presented in [21].

The structure of this paper is as follows: Section 2 describes the different sources of evidence we use and analyzes their behaviour. Section 3 presents two simple ways of combining the different sources together yielding better results than with the single ones. Before concluding, we discuss some related work in Section 4.

2 Heterogeneous Evidence

In this section, we describe the different sources of evidence we use in our approach. What we mean by 'source of evidence' is the implicit knowledge about the validity of a certain semantic relation, i.e. the *is-a* relations we consider here, in unstructured resources such as corpora, the World Wide Web etc. Before describing the particular sources of evidence we use in our approach, we would like to motivate the necessity of using different sources of evidence. Certainly, the learning of taxonomic relations can be seen as a classification task. In fact, given two terms, say *conference* and *event*, they could either be related in three different ways: is-a(conference,event), is-a(event,conference), siblings(conference,event), or they could also be taxonomically unrelated. So, given a certain set of terms, if we are able to assign each pair of terms the correct relation out of the above ones, we will yield a correct and complete concept hierarchy for these terms. Thus the learning of taxonomic relations can in fact be seen as a classification task with a special category for 'unrelatedness'. In this line, as a first step towards learning complete concept hierarchies, in this paper we focus on an easier classification task, i.e. given two terms we will decide if they stand in a *is-a*-relation or not. Considering this task, it is thus intuitive to gather as many different sources of evidence as possible and choose the relation with maximal evidence with regard to all these different sources. In what follows we describe the sources of evidence we use and illustrate them with a running example centered around the *conference* concept.

As underlying corpus for the corpus-based approaches 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 ontology we consider for evaluating our approach is the tourism reference ontology in [21] which was modeled by an experienced ontology engineer. 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 non-direct relations. It is important also to mention that we consider only pairs of terms/concepts contained in the ontology, which we thus aim at 'reproducing' with our approach.

2.1 Hearst Patterns

The first source of evidence we consider are lexico-syntactic patterns matched against a certain corpus in line with the approaches of [14], [5] and [26]. In particular, the patterns we use are taken from [14][1]:

- NP_0 such as NP_1, NP_2, ..., NP_{n-1} (and|or) NP_n
- such NP_0 as NP_1, NP_2, ... NP_{n-1} (and|or) NP_n
- NP_1, NP_2, ..., NP_n (and|or) other NP_0
- NP_0, (including|especially) NP_1, NP_2, ..., NP_{n-1} (and|or) NP_n

According to Hearst, from the above patterns we can derive that for all $NP_i,\ 1 \leq i \leq n,\ isa(head(NP_i), head(NP_0))$[2]. Now given two terms t_1 and t_2 we record how many times a Hearst-pattern indicating an *is-a*-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 POS-tags to match NP's. In particular, we use the tagger described in [31] 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 is-a relations, where the number in brackets gives the normalized value as described above:

[1] NP_i stands for a noun phrase.

[2] Actually [14] 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.

is-a$_{HEARST}$(conference,event) (0.44)
is-a$_{HEARST}$(conference,body) (0.22)
is-a$_{HEARST}$(conference,meeting) (0.11)
is-a$_{HEARST}$(conference,course) (0.11)
is-a$_{HEARST}$(conference,activity) (0.11)

The first interesting observation here is that, despite of using quite a big corpus, Hearst patterns appear relatively rarely. In fact, in the whole corpus we only matched 8647 Hearst patterns, of which only 99 correspond to pairs of concepts in our tourism ontology. Figure 1 shows the behaviour in terms of Recall, Precision and F-Measure on our dataset (the tourism ontology). In particular, we use a threshold parameter t and only consider the results for which the normalized value is above the threshold. In the following figures we omit the values for $t = 1$ as the precision P is always 1, the recall R is 0 and thus the F-Measure 0. The results of Hearst's approach show in fact that we can get quite high precisions depending on the threshold used. However, it becomes also clear that the recall of Hearst's approach is very low as the patterns appear rarely. The top F-Measure for the Hearst approach is F=6.92% ($t = [0 : 0.1]$), corresponding to a Precision of 29.1% and a Recall of 3.93%. The top precision is 39.02% ($t = [0.5 : 0.9]$) at a recall of 2.52%. Thus in terms of precision the results are lower as for example the ones cited in [14], but in contrast to there we are considering a given domain and evaluating with regard to a given concept hierarchy.

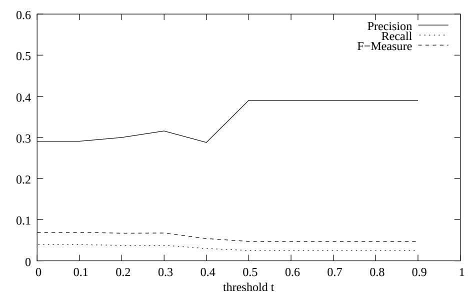

Figure 1. Results Hearst

2.2 WordNet

As a further source of evidence we use the hypernymy information from WordNet[3]. Actually, WordNet can not be seen as an unstructured source of evidence, but the information 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 mention 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 $max(\frac{|paths(senses(t_1), senses(t_2))|}{|senses(t_1)|}, 1)$. 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:

[3] We used version 1.7.1 for our experiments.

is-a$_{WN}$(conference, organization) (1)
is-a$_{WN}$(conference,group) (0.67)

Besides the above strategy for integrating information from WordNet, we also considered a variant in which we considered all the senses of t_1 and one in which we considered only the first, i.e. most frequent, sense of t_1 as specified by the formula $max(|paths(first\_sense(t_1), senses(t_2))|, 1)$. This value is obviously 0 or 1. The results in terms of precision, recall and F-measure over threshold t are given in figures 2 and 3, respectively.

The precision for the is-a 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. Obviously, taking only the first sense in WordNet helps in reducing the ambiguity and thus increasing the precision (compare figure 3). When considering all senses the best F-Measure is F=10.84% at $t = 0.2$ with a Precision of P=21.60% and a Recall of R=7.23%. When considering only the first sense the best F-Measure is F=8.88% with a Precision of P=30.55% and a Recall of R=5.20%. So, the precision has increased at the cost of a lower recall yielding overall a lower F-Measure. It is certainly striking that the precision of the relations found in WordNet is so low. It is important to notice 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*... This raises of course the question in how far automatically learned ontologies can be actually evaluated against a certain gold standard. Another interesting alternative would thus be to involve humans in the evaluation accepting or rejecting a certain relation proposed by the system.

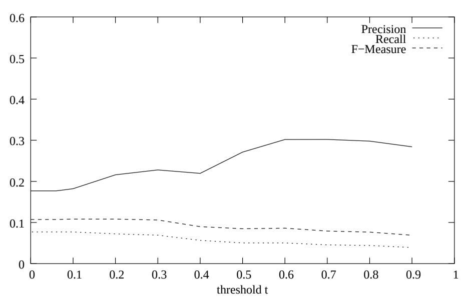

Figure 2. Results WordNet

2.3 'Vertical Relations'-Heuristic

In order to identify further is-a relations, we make use of a heuristic used by [33] which we will henceforth call *'vertical relation'*-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 is-a(t_1,t_2). As an example, according to this heuristic, we might derive that t_1='international conference' and t_2='conference'

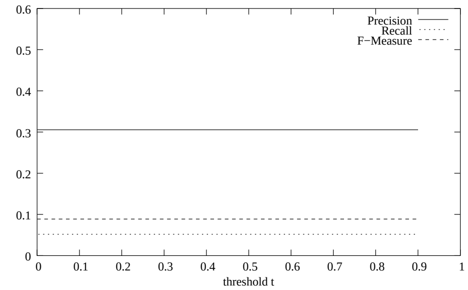

Figure 3. Results WordNet (most frequent sense)

are related by an is-a relation, i.e. is-a$_{HEURISTIC}$(international conference,conference). Figure 4 shows the results of this heuristic in terms of Precision, Recall and F-Measure over the threshold t. Obviously, in this case these values are constant as the heuristic can be either true (1) or false (0). The best F-Measure here was F=7.02%, corresponding to a Precision of P=50% and a Recall of R=3.77%. So as in the case of the Hearst patterns the precision seems reasonable while the recall is very low.

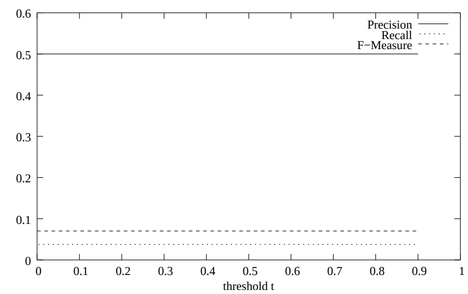

Figure 4. Results Heuristic

2.4 World Wide Web

Certainly, when using a corpus we have to cope with typical data sparseness problems [34]. On the other hand, some researchers have shown that the World Wide Web is an attractive way of reducing data sparseness [20, 11, 29]. In this line, following [22, 6], we use the Google API[4] to count the matches of a certain pattern 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:

$< t_1 >$s such as $< t_2 >$
such $< t_1 >$s as $< t_2 >$
$< t_1 >$s, including $< t_2 >$
$< t_1 >$s, especially $< t_2 >$
$< t_2 >$ and other $< t_1 >$s
$< t_2 >$ or other $< t_1 >$s

As in [14], these patterns are indicators for a corresponding taxonomic relation is-a$_{WWW}$(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*

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

concept and other terms in the ontology we consider; the number in parenthesis indicates the normalized number of hits returned by the Google API:

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

It is important to note that due to the patterns we use, we get no information for nouns which do not form their plural regularly, e.g. *activity*. This is certainly a problem with this approach. The results of the WWW approach in terms of Precision, Recall and F-Measure over the threshold t are given in Figure 5. The best F-Measure here was F=18.85% at $t = 0.04$ with a precision of P=15.77% and a recall of R=23.43%. 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.

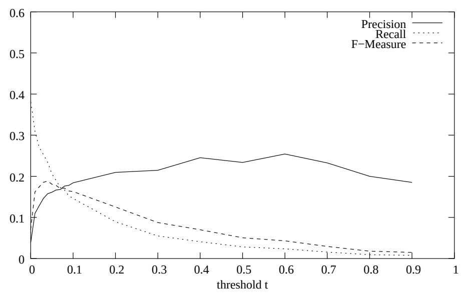

Figure 5. Results WWW

3 Combining Heterogeneous Evidence

In order to combine the four approaches described above, we explored two very simple strategies, one in which we computed the mean of the value returned by the above approaches and one in which we took the maximum value of the four approaches[5]. Take for example the results of the different approaches for the *conference* concept given in Table 1. Table 2 gives the values for these pairs according to both strategies. The results of the *mean* strategy are given in Figure 6 for the version considering all the WordNet senses for a word and in 7 for the version considering only the first sense of a word. The *mean* strategy with all senses yielded an F-Measure of F=20.84% (P=17.16%, R=29.84%) at $t = 0.01$ and the one with the first senses one of F=21.8% (P=17.38%, R=29.24%) at $t = 0.01$. The *max* strategy with all senses (see figure 8) yielded an F-Measure of F=20.87% ($t = 0.04$), corresponding to a precision and recall of P=16.03% and R=29.87%, respectively. The one only considering the first sense (compare figure 9) yielded F=21.81% ($t = 0.04$) with a precision of P=17.38% and a recall of R=29.25%. Thus, the best results of both strategies are comparable and clearly improve the best result of the WWW approach. In general, while the recall appears to be reasonable, the precision is still quite low for the best results in terms of

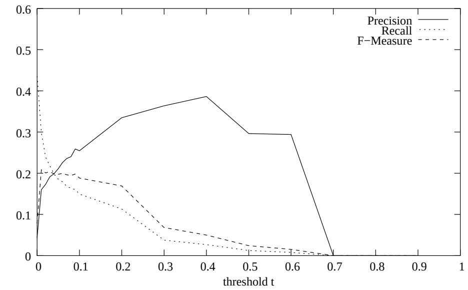

Figure 6. Mean Strategy (all senses)

the F-Measure. However, as Figure 7 shows, we can get quite high precisions (at a much lower recall) when using higher thresholds. The choice of the threshold thus clearly depends on the application in question.

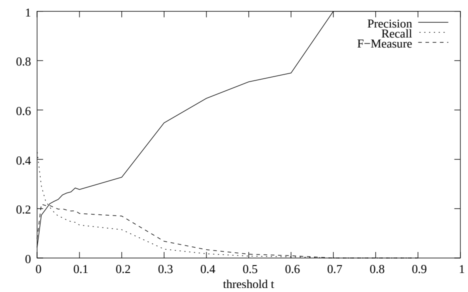

Figure 7. Mean Strategy (first sense)

3.1 Discussion

Our strategies to combine the different approaches together are certainly very simple but we have demonstrated that they lead to better results thus showing that there is a lot of potential in the combination of different approaches for the purpose of learning is-a relations. The main challenge here certainly is to find an optimal combination. We believe that machine learning can offer a solution to this, but this means that we will have to cope somehow with the large amount of negative examples. As already mentioned above, this remains challenging work for the future. The main contribution of this paper is thus a systematic analysis of some state-of-the-art approaches to learning taxonomic relations as well as the claim that combining different approaches can definitely lead to improved results.

4 Related Work

In this section, we discuss some work related to the automatic acquisition of taxonomies. The main paradigms for learning taxonomic relations exploited in the literature are on the one hand clustering approaches based on the distributional hypothesis ([13]) and on the other hand approaches based on matching lexico-syntactic patterns which convey a certain relation in a corpus.
One of the first works on clustering nouns was the one by Hindle ([15])[6], in which nouns are grouped into classes according to the extent to which they appear in similar verb frames. In particular, he takes into account nouns appearing as subjects and objects of verbs, but does not distinguish them in his similarity measure. Pereira et al. ([25]) also present a top-down clustering approach to build an unlabeled hierarchy of nouns. The work of Faure et al. ([9]) is also based

[5] In fact, we also explored different machine learning algorithms from WEKA such as C4.5 decision trees, support vector machines, Naive Bayes, a Perceptron and a Multi-layer Perceptron but could not manage to learn a reasonable model due to the huge number of negative examples. This problem is shared by other approaches within NLP such as coreference resolution ([24]). Coping with this problem in fact remains for further work.

[6] One of the reviewers draw our attention to the even earlier work in [17].

(t_1,t_2)	HEARST	WN (all senses)	WN (first sense)	HEURISTIC	WWW
(conference,event)	0.44	0	0	0	0.25
(conference,organization)	0	1	0	0	0.05
(conference,group)	0	0.33	1	0	0.02
(conference,service)	0	0	0	0	0.27
(conference,area)	0	0	0	0	0.11
(conference,information)	0	0	0	0	0.04

Table 1. The *conference* example: values of the different approaches

(t_1,t_2)	Mean (all senses)	Mean (first sense)	Max (all senses)	Max (first sense)
(conference,event)	0.17	0.17	0.44	0.44
(conference,organization)	0.26	0.01	1	0.05
(conference,group)	0.09	0.26	0.33	1
(conference,service)	0.07	0.07	0.27	0.27
(conference,area)	0.03	0.03	0.11	0.11
(conference,information)	0.01	0.01	0.04	0.04

Table 2. The *conference* example: results of the combination strategies

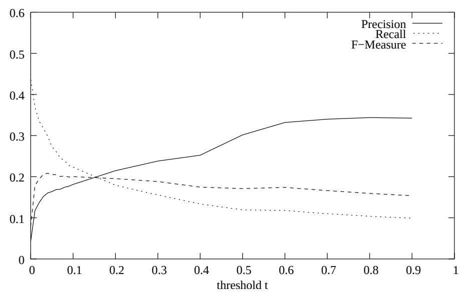

Figure 8. Max strategy (all senses)

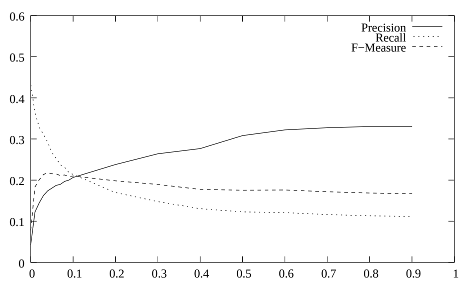

Figure 9. Max strategy (fi rst sense)

on the distributional hypothesis; they present an iterative bottom-up clustering approach of nouns appearing in similar contexts. In each step, they cluster the two most similar extents of some argument position of two verbs. However, their approach requires manual validation after each clustering step so that in our view it can not be called *unsupervised* or *automatic* anymore. Caraballo ([4]) also uses clustering methods to derive an unlabeled hierarchy of nouns by using data about conjunctions of nouns and appositive constructs collected from the Wall Street Journal corpus. Interestingly, at a second step she also labels the abstract concepts of the hierarchy by considering the Hearst patterns in which the children of the concept in question appear as hyponyms. The most frequent hypernym is then chosen in order to label the concept. At a further step she also compresses the produced ontological tree by eliminating internal nodes without a label. The final ontological tree is then evaluated by presenting a random choice of clusters and the corresponding hypernym to three human judges for validation. Bisson et al. ([3]) present an interesting framework and a corresponding workbench - Mo'K - allowing users to design conceptual clustering methods to assist them in an ontology building task. In particular they use bottom-up clustering and compare different similarity/distance metrics as well as different pruning parameters. In more recent work, viz. [7], also Formal Concept Analysis has been applied as a conceptual clustering method to derive concept hierarchies from text.

Concerning the use of lexico-syntactic patterns denoting a certain semantic relation we are aware of the approaches in [14], [5] and [26]. Hearst [14] aims at the acquisition of hyponym relations from Grolier's American Academic Encyclopedia. In order to identify these relations, she makes use of lexico-syntactic patterns manually acquired from her corpus. Similar in spirit to Hearst's approach are the ones presented in [5] for learning part-of relations, [26], in which the focus is to learn different relations for an anaphora resolution task as well as [2, 19, 32]. The approaches of Hearst and others other are characterized by a (relatively) high precision in the sense that the quality of the learned relations is very high. However, these approaches suffer from a very low recall which is due to the fact that the patterns are very rare.

In order to overcome such data sparseness problems, in our approach we resort to the WWW as for example in [22], where Hearst patterns are also searched in the WWW by using the Google API in order to acquire background knowledge for anaphora resolution, as well as in [1], where related texts are crawled from the Web to enrich a given ontology. In [6] a similar approach was employed to find the best concept for an unknown instance in a given ontology.

Other related approaches are the ones in [33] from which we reused the 'vertical relations'-heuristic. Another interesting approach is the one in [30], where a hierarchy between nouns is derived automatically by considering the document a certain term appears in as context. In particular, they 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.

5 Conclusion and Further Work

We have presented a novel approach to learning concept hierarchies from different and heterogeneous sources of evidence. In particular, we have considered certain lexico-syntactic patterns matched in a corpus as well as in the World Wide Web, information from WordNet as well as the 'vertical relations' heuristic presented in [33]. We have systematically analyzed these different approaches in terms of Precision, Recall and F-Measure and showed that a simple combination strategy already improves the results.

It remains further work to find out if other sources of evidence could be integrated into our approach. Moreover, we have neglected at all the context of the domain we want to acquire an ontology for, i.e. in our case the tourism domain. In fact, all the sources of evidence we consider are of a very general nature. In this sense it could turn out to be useful to only consider domain-specific text collections instead of a general corpus such as the BNC and to consider only pages in the World Wide Web related to the domain in question as in [1].

More importantly, we insist that it remains as a challenge to determine the optimal strategy to combine the different approaches. In order to apply machine learning techniques for this purpose, we will need to cope with the high number of negative examples, which is a non-trivial problem as discussed in [27]. In addition, it seems necessary to explore alternative evaluation methodologies in which humans are involved in the cycle assessing the quality of the learned relations. Otherwise when evaluating against a certain standard, the system is penalized for finding relations which are correct or at least possible but which are not in the target ontology.

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 valuable discussions. Thanks in particular to Andreas Hotho, Marko Grobelnik, Johanna Völker as well as the reviewers of the workshop for feedback and comments on earlier versions of this paper.

REFERENCES

[1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, 'Enriching very large ontologies using the WWW', in *Proceedings of the Workshop on Ontology Construction of the ECAI*, (2000).

[2] K. Ahmid, 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)*, (2003).

[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*, (2000).

[4] 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*, pp. 120–126, (1999).

[5] E. Charniak and M. Berland, 'Finding parts in very large corpora', in *Proceedings of the 37th Annual Meeting of the ACL*, pp. 57–64, (1999).

[6] P. Cimiano, S. Handschuh, and S. Staab, 'Towards the self-annotating web', in *Proceedings of the 13th World Wide Web Conference*, (2004).

[7] P. Cimiano, A. Hotho, and S. Staab, 'Clustering ontologies from text', in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, (2004).

[8] 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)*, (2001).

[9] D. Faure and C. Nedellec, 'A corpus-based conceptual clustering method for verb frames and ontology', in *Proceedings of the LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, ed., P. Velardi, (1998).

[10] C. Fellbaum, *WordNet, an electronic lexical database*, MIT Press, 1998.

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

[12] 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*, (1998).

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

[14] M.A. Hearst, 'Automatic acquisition of hyponyms from large text corpora', in *Proceedings of the 14th International Conference on Computational Linguistics*, (1992).

[15] D. Hindle, 'Noun classifi cation from predicate-argument structures', in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 268–275, (1990).

[16] L. Hirschman and N. Chinchor, 'Muc-7 named entity task defi nition', in *Proceedings of the 7th Message Understanding Conference (MUC-7)*, (1997).

[17] L. Hirschman, R. Grishman, and N. Sager, 'Grammatically-based automatic word class formation', *Information Processing and Management*, **11**, (1975).

[18] A. Hotho, S. Staab, and G. Stumme, 'Ontologies improve text document clustering', in *Proc. of the ICDM 03, The 2003 IEEE International Conference on Data Mining*, pp. 541–544, (2003).

[19] C. Jouis, *Contribution a la conceptualisation et a la Modelisation des connaissances a partir d'un analyse linguistique de textes. Realisation d'un prototype: le systeme SEEK*, Ph.D. dissertation, Universite Paris III - Sorbonne Nouvelle, 1993.

[20] F. Keller, M. Lapata, and O. Ourioupina, 'Using the web to overcome data sparseness', in *Proceedings of EMNLP-02*, pp. 230–237, (2002).

[21] A. Maedche and S. Staab, 'Measuring similarity between ontologies', in *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW)*. Springer, (2002).

[22] K. Markert, N. Modjeska, and M. Nissim, 'Using the web for nominal anaphora resolution', in *EACL Workshop on the Computational Treatment of Anaphora*, (2003).

[23] D. Moldovan and R. Mihalcea, 'Using wordnet and lexical operators to improve internet searches', *IEEE Internet Computing*, **4**(1), (2000).

[24] Vincent Ng and Claire Cardie, 'Combining sample selection and error-driven pruning for machine learning of coreference rules', in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, (2002).

[25] 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*, pp. 183–190, (1993).

[26] 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).

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

[28] P. Resnik, 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *Journal of Artificial Intelligence Research*, **11**, 95–130, (1999).

[29] P. Resnik and N. Smith, 'The web as a parallel corpus', *Computational Lingusitics*, **29**(3), (2003).

[30] M. Sanderson and B. Croft, 'Deriving concept hierarchies from text', in *Research and Development in Information Retrieval*, 206–213, (1999).

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

[32] P. Seguela, *Construction de modeles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*, Ph.D. dissertation, Universite de Toulousse, 2001.

[33] P. Velardi, P. Fabriani, and M. Missikoff, 'Using text processing techniques to automatically enrich a domain ontology', in *Proceedings of the ACM International Conference on Formal Ontology in Information Systems*, (2001).

[34] G. Zipf, *Selective Studies and the Principle of Relative Frequency in Language*, Cambridge, 1932.