Finding the Appropriate Generalization Level for Binary Ontological Relations Extracted from the Genia Corpus

P. Cimiano[1], M. Hartung[1], E. Ratsch[2]

[1]Institute AIFB, University of Karlsruhe, cimiano@aifb.uni-karlsruhe.de, hartung@urz.uni-heidelberg.de
[2] Bioinformatics, University of Würzburg, Esther.Ratsch@biozentrum.uni-wuerzburg.de

Abstract

Recent work has aimed at discovering ontological relations from text corpora. Most approaches are based on the assumption that verbs typically indicate semantic relations between concepts. However, the problem of finding the appropriate generalization level for the verb's arguments with respect to a given taxonomy has not received much attention in the ontology learning community. In this paper, we address the issue of determining the appropriate level of abstraction for binary relations extracted from a corpus with respect to a given concept hierarchy. For this purpose, we reuse techniques from the subcategorization and selectional restrictions acquisition communities. The contribution of our work lies in the systematic analysis of three different measures. We conduct our experiments on the Genia corpus and the Genia ontology and evaluate the different measures by comparing the results of our approach with a gold standard provided by one of the authors, a biologist.

1. Introduction

A lot of effort has been devoted to discovering ontological relations from text corpora in recent years (Mädche and Staab, 2000; Yamaguchi, 2001; Kavalec and Svátek, 2005; Ciaramita et al., 2005; Schutz and Buitelaar, 2005). Relations together with ontological restrictions on their arguments are needed for many applications, especially in the field of natural language processing. Ontological restrictions can, for example, be used as a basis to capture the selectional restrictions and preferences of verbs for disambiguation purposes. Relations as well as inference rules defined on their basis have important applications in question answering (Lin and Pantel, 2001). Further, relations automatically derived from a corpus can assist a domain expert in ontology engineering.

Most approaches to learning ontological relations from text are based on the assumption that verbs typically indicate semantic relations between concepts, e.g. (Kavalec and Svátek, 2005; Ciaramita et al., 2005; Schutz and Buitelaar, 2005). However, the problem of finding the appropriate generalization level for the verb's arguments with respect to a given taxonomy has not received much attention in the knowledge acquisition community. In fact, the only works we are aware of along these lines are the ones in (Faure and Nedellec, 1998; Mädche and Staab, 2000; Ciaramita et al., 2005). A very related problem is the acquisition of selectional restrictions (compare (Ribas, 1995; Resnik, 1997; Clark and Weir, 2002)). In this paper we address the issue of determining the appropriate level of abstraction for binary relations extracted form a corpus with respect to a given concept hierarchy. For this purpose, as in (Ciaramita et al., 2005), we reuse techniques from the subcategorization and selectional restrictions acquisition communities. We conduct our experiments on the Genia corpus and the Genia ontology[1]. The contribution of our work lies in the systematic analysis of three different measures. We evaluate the different measures by comparing the results of our approach with a gold standard provided by the third author, a biologist.

2. Approach

In our approach, verb frames are extracted using Steven Abney's chunker CASS (Abney, 1996). From CASS's output, we extract tuples NP-V-NP and NP-V-P-NP. We construct binary relations from these tuples, using the lemmatized verb V (with the preposition P if applicable) as corresponding relation label and the head of the NP phrases as concepts for the domain and range of the relation. In particular, we only consider nouns as concepts which also appear in the Genia ontology. Our aim is then to find the most general and appropriate concept for the domain and range of the relation on the basis of the different examples found in the corpus. For illustration purposes, let us consider the input sentences marked with (a) and the CASS output in (b):

(1) a. This bipartite motif consists of an N-terminal POU-specific domain.
b. consist(subj:bipartite motif, of: N-terminal POU-specific domain)

(2) a. Infection leads to severe acute disease in macaques.
b. lead(subj:infection, to:disease, in: macaque)

(3) a. Lipoarabinomannan releases IL-6 in a dose-response manner.
b. release(subj:Lipoarabinomannan, obj:IL-6, in:dose-response manner)

While the NP-V-NP pattern can be generally mapped to Subj-V-Obj structures without producing too many errors, the NP-V-P-NP pattern generates substantial noise due to PP-attachment ambiguities. Particularly, CASS does not differentiate between PPs functioning as oblique arguments of the verb (as in (1) and (2)) and facultative adjuncts (as in (3)). However, we decided to keep this pattern and assume that every PP attaches to the preceding verb. For each of these patterns, we then create binary relations labeled with

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/

the verb V (and the preposition P if applicable), relying on the semantic annotations of the Genia corpus to map the arguments to corresponding concepts for the domain and range of the relation. Tuples which would be extracted from the CASS output above are for example:

consist_of(motif, domain)
lead_to(infection, disease)
release(Lipoarabinomannan, IL-6)
release_in(Lipoarabinomannan, dose-response manner)

It is important to emphasize that we rely on the semantic annotation in the Genia corpus to map the verbs' arguments to concepts in the ontology.

2.1. Generalizing Verb Frames

Having thus collected a number of labeled relations from the corpus, our aim is to find the most appropriate generalization for the concepts within the domain and the range of each relation on the basis of the different examples found in the corpus. For this purpose, we experiment with three different measures:

- the conditional probability of a concept given a verb slot,
- the pointwise mutual information between a concept and a verb slot,
- a χ^2-based measure.

We briefly describe the three measures in the following section and illustrate them on the basis of an example.

2.2. Measures

As an illustrating example, let us consider the object position of the verb *activate*. Let us further assume that the objects appearing in the corpus for *activate* together with their frequencies are the following:

protein_molecule:	5
protein_family_or_group:	10
amino_acid:	10

The above example reflects the empirically observed frequencies of concepts in the respective argument position before the propagation of frequencies along the taxonomy, i.e. the hierarchical structure of the Genia ontology is not taken into account. In order to find the appropriate concept for a certain slot with respect to the hierarchy, we examine three measures which are described in the following and illustrated according to this example.

2.2.1. Conditional Probability

The first method examined calculates for a certain slot s of a verb v the conditional probability that a concept c appears in this slot, propagating the frequencies along the concept hierarchy (see Figure 1), and then chooses the concept maximizing this value:

$$c_{v_s} := argmax_c \; P(c|v_s)$$

If there are several concepts with the same value, we choose the most specific ones, leaving out the concepts which subsume them. For our example we would get:

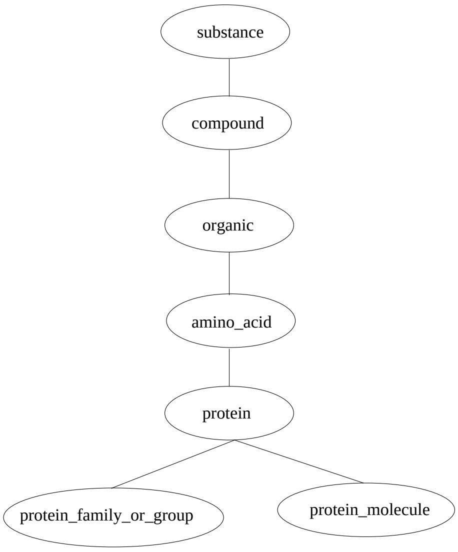

Figure 1: Part of the Genia ontology

P(protein_molecule | activate_obj) = $\frac{5}{25}$ = 0.2
P(protein_family_or_group | activate_obj) = $\frac{10}{25}$ = 0.4
P(protein | activate_obj) = $\frac{15}{25}$ = 0.6
P(amino_acid | activate_obj) = $\frac{25}{25}$ = 1
P(organic | activate_obj) 1
P(compound | activate_obj) = 1
P(substance | activate_obj) = 1

So we would choose *amino_acid* as the appropriate generalization for the object position of *activate*.

2.2.2. Pointwise Mutual Information

The method based on the mutual information penalizes the conditional probability value above in case the concept c is very frequent. The underlying hypothesis is that a concept occurring very frequently in the context of all verbs is not a good generalization candidate for a specific verb. The best concept is determined by the following formula:

$$\begin{aligned} c_{v_s} &= argmax_c \; PMI(c, v_s) \\ &= argmax_c \; log_2 \frac{P(c|v_s)}{P(c)} \end{aligned}$$

Now assuming a probability $P(amino\_acid) = \frac{825}{3050} = 0.27$ for *amino_acid* occurring as the object of *activate* and $P(protein) = \frac{415}{3050} = 0.14$ for *protein* (compare Tables 1 and 2), we would get:

PMI(protein|activate_obj) = $log_2 \frac{0.6}{0.14} = 2.1$
PMI(amino_acid|activate_obj) = $log_2 \frac{1}{0.27} = 1.89$

According to the PMI-measure, we would thus choose *protein* as the most appropriate generalization.

2.2.3. A χ^2-based measure

The measure based on the χ^2-test substantially differs from the other measures in the sense that it does not compare conditional probabilities but contingencies between two

Table 1: 2-by-2 χ^2 table for protein as range of activate

	range(activate)	range(¬ activate)
protein	15	400
¬ protein	35	2600

Table 2: 2-by-2 χ^2 table for amino_acid as range of activate

	range(activate)	range(¬ activate)
amino_acid	25	800
¬ amino_acid	25	2200

variables. The procedure performs a test whether the two variables are statistically independent or not. We apply χ^2 as proposed in (Clark and Weir, 2002), testing the contingencies between v_s and the concept c as well as its possible generalizations $c'_1, ...c'_n$ in an iterative manner. The assumption is that we can generalize c to c'_i as long as χ^2 reveals v_s and c'_i to be statistically dependent. A result is considered significant with regard to a significance level $\alpha = 0.05$ if the χ^2 value within our 2 × 2 χ^2-matrix exceeds the typically assumed critical value of 3.84.

The formula used for the χ^2 test is:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} are the so called *observed frequencies* as calculated on the basis of the corpus and given in row i and column j in Tables 1 and 2 and E_{ij} are the expected frequencies calculated under the assumption of independence between v_s and c'_i.

For the 2×2 case we have (compare (Manning and Schütze, 1999)):

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

where N is the sum of all the frequencies in the table. For the examples in Table 1 and 2 we thus yield:

$$\chi^2(range(activate), protein) =$$
$$= \frac{3050(15 * 2600 - 400 * 35)^2}{415 \times 50 \times 3000 \times 2635} = 11.62$$
$$\chi^2(range(activate), amino\_acid) =$$
$$= \frac{3050(25 * 2200 - 800 * 25)^2}{825 \times 50 \times 3000 \times 2225} = 13.57$$

Thus, in both cases we get a significant result at a level of $\alpha = 0.05$. The generalization from *protein* to *amino_acid* is thus a valid one according to the χ^2-based measure. The variations in the predicted concept for the range of *activate* show that the measure chosen can indeed have a decisive impact on the results.

3. Evaluation

In order to evaluate the different measures we propose, we applied our preprocessing to the Genia corpus (Ohta et al., 2002). Overall, the corpus contains 18.546 sentences with 509.487 words and 51.170 verbs. We use the semantic annotations of the Genia corpus to map the subject and object of verb phrases to the Genia ontology. The domain and range of the extracted relations are then generalized with respect to the Genia ontology using the measures described above. For the evaluation of the different measures, one of the authors, a biologist, specified the ideal domain and range for 100 binary relations corresponding to the 100 most frequent patterns extracted with the approach based on CASS as described above. The average frequency of occurrence for the verbs of these 100 patterns is around 17.51, with a minimum of 3 and a maximum of 148 occurrences. Of these 100 relations, 15 were regarded as inappropriate by our evaluator, such that the evaluation is carried out on the remaining 85 relations.

Our biologist specified a number of concepts from the Genia ontology as the best generalization for the domain and range of each relation denoted by the verb. In some cases, she was also able to specify one single 'best concept' out of several possible candidates. In general, however, she specified a set of concepts generalizing each argument position. The output of our approach is compared with this gold standard using the different measures described above in terms of:

- direct matches for domain and range (DM),
- average distance in terms of number of edges in the taxonomy between correct and predicted concept (AD), and
- a symmetric variant of the Learning Accuracy (LA) defined in (Hahn and Schnattinger, 1998).

The different measures are formalized in Figure 2. There, R denotes the set of relations in the output of our system. Further, for $r \in R$ we define $dom_S(r)$ as the domain produced by our system and $dom_G(r)$ as the domain as specified in the gold standard; $range_S(r)$ and $range_G(r)$ are defined analogously. Please note that these functions all return sets. The system returns more than one concept in case there is a tie, and our annotator used more than one concept in most cases, indicating the most appropriate wherever possible.

The learning accuracy LA is inspired by the corresponding measure introduced in (Hahn and Schnattinger, 1998). However, we consider a slightly different formulation of the learning accuracy as defined in (Mädche and Staab, 2000). The measure of Hahn and our learning accuracy measure are not totally equivalent. The main difference is that we measure the distance between nodes in terms of edges – instead of nodes as in Hahn's version – and we do not need any case distinction considering whether the classification was correct or not. Additionally, in contrast to Hahn's learning accuracy, our measure is symmetric. The learning accuracy between two concepts is defined as:

$$LA(a, b) := \frac{\delta(top, c) + 1}{\delta(top, c) + \delta(a, c) + \delta(b, c) + 1}$$

where $c = lcs(a,b)$, i.e. c is the least common subsumer of a and b in the taxonomy and δ measures the distance between two nodes as the number of edges between them. In particular, the distance is defined as following:

$$\delta(a,b) := \delta(a, lcs(a,b)) + \delta(b, lcs(a,b))$$

where δ measures the distance in terms of edges and obviously $\delta(a,a) = 0$.

Due to the fact that our system as well as the annotator specified a set of possible concepts as domain and range of the relations, we decided to consider three evaluation modes: i) *optimistic*, ii) *average*, and iii) *pessimistic*. The *optimistic* version compares that concept our system predicts for a certain position of a relation with the concept in the gold standard yielding the best result with respect to the given evaluation measure. The *pessimistic* version chooses the concepts in the output of the system and the gold standard yielding the worst measures, whereas the *average* averages the results of the evaluation measures for all combinations of concepts in the system's output and the gold standard. Table 3 summarizes our results. It shows, for each measure, the percentage of direct matches, as well as the optimistic, average and pessimistic variants of the average distance and learning accuracy. The main conclusion is that the conditional probability consistently outperforms all other measures with respect to all evaluation modes.

4. Discussion and Related Work

Our results show that the conditional probability is a reasonable measure to find the correct level of generalization with respect to a given concept hierarchy for verb-based relations extracted from a (semantically annotated) corpus. The conditional probability outperforms the χ^2 based measure in terms of direct matches, average distance and learning accuracy, which in turn outperforms the pointwise mutual similarity measure. An important observation is that in many cases our human evaluator has chosen abstract concepts, which are in general disfavored by the PMI-measure. This explains why the PMI measure performs so poorly. Our approach is similar to the work of Resnik (Resnik, 1997) and Ribas (Ribas, 1995) on acquiring selectional restrictions. Both have formulated the problem of finding the right level with respect to WordNet as the one of finding the maximum with respect to a given statistical measure. Resnik examines a measure called *association strength*, which takes into account the *selectional strength* of a verb, i.e. the Kullback-Leibler divergence between the prior and posterior distributions of a noun and a verb slot. Ribas examines a variety of measures and, as in our case, concludes that the PMI and the χ^2 measure do not perform as well as the other measures. McCarthy (McCarthy, 1997) presents an approach based on the Minimum Description Length (MDL) principle originally developed by Li and Abe (Li and Abe, 1998). All the above approaches evaluate their models on word sense disambiguation tasks and are thus not directly comparable to the results of our approach. Further, our approach relies on the semantic annotations of the Genia corpus, such that we are not faced with sense ambiguity as the above approaches.

Recently, Ciaramita et al. (Ciaramita et al., 2005) have applied a variant of the model for acquisition of selectional restrictions of Clark and Weir (Clark and Weir, 2002) to the Genia corpus. The authors rely on the approach of (Clark and Weir, 2002) to determine whether using a hypernym instead of the hyponym leads to significantly different probabilities. They compare the probability $p(r|c,s)$ with $p(r|c',s)$ where c' is a superconcept of c. If $p(r|c',s)$ and $p(r|c,s)$ do not significantly differ, c' is regarded as an appropriate generalization. The authors present a twofold evaluation of their approach. On the one hand, they present the learned relations to a biologist – actually the same as in our case – for manual validation, coming to the conclusion that 83.3% of the learned relations are correct, and furthermore 53.1% of the generalized relations have been generalized appropriately. Mädche and Staab (Mädche and Staab, 2000) present an approach relying on an algorithm for mining generalized association rules to find conceptual relations between words at the appropriate level of abstraction with respect to a given taxonomy. In their approach, transactions are defined in terms of words occurring together in certain syntactic dependencies. Generalization of argument positions is achieved by removing those association rules subsumed by some other association rule. Mädche and Staab achieve a best recall and precision of R=13% and P=11% in terms of direct matches with respect to the gold standard. The method of Yamaguchi (Yamaguchi, 2001) essentially implements word space (Schütze, 1993) and assumes that there is a relation between words which are similar beyond a certain threshold. Yamaguchi states that out of 90 extracted concept pairs, 53 are 'advisable'. This result can be regarded as corresponding to a precision of about 59%. However, Yamaguchi does not address the problem of finding the right level of abstraction and does not derive labeled, but 'anonymous' relations. Schutz and Buitelaar (Schutz and Buitelaar, 2005) apply shallow linguistic analysis to extract concept–verb–concept triples and filter these on the basis of a χ^2-based measure. They evaluate their approach in terms of recall and precision with respect to a gold standard, achieving a precision between 9.1% and 11.9%, depending on the evaluation set used. In general, it is important to emphasize that there is a substantial difference between *a priori* and *a posteriori* evaluations. In *a priori* evaluations, the gold standard is constructed independently of the results of the system, and the system is then evaluated with respect to the gold standard in a strict way. In *a posteriori* evaluations, the results of a system are presented to the evaluator, who then classifies the results of the system. In the first case, the system can be penalized still if its results are reasonable and just because an answer diverges from the one in the gold standard. *A posteriori* evaluation differs in this respect as the results merely depend on how inclined the evaluator is to regard the suggestions of the system as correct. The difference between *a priori* and *a posteriori* evaluation is illustrated by Schutz and Buitelaar, who present their results both in terms of *a priori* as well as *a posteriori* evaluation. With respect to the *a posteriori* evaluation, they report an average precision between 17.7% and 23.9%, yielding approx. 10% higher results compared to the *a priori evaluation*. Examples for

$$\begin{aligned} DM &= \frac{\textit{direct matches for domain} + \textit{direct matches for range}}{2\ |R|} \\ AD &= \frac{\sum_{r \in R} \delta(dom_S(r), dom_G(r)) + \delta(range_S(r), range_G(r))}{2\ |R|} \\ LA &= \frac{\sum_{r \in R} LA(dom_S(r), dom_G(r)) + LA(range_S(r), range_G(r))}{2\ |R|} \end{aligned}$$

Figure 2: Evaluation Measures

Table 3: Results for the different measures

	DM	AD			LA		
		opt.	avg.	pess.	opt.	avg.	pess.
Conditional	33.53%	1.21	1.76	2.22	70.40%	60.57%	53.24%
PMI	13.53%	3.28	3.76	4.19	48.65%	43.06%	38.62%
χ^2	26.79%	2.63	3.44	4.15	56.71%	46.19%	38.48%

a priori evaluations are the ones of Mädche et al., Schutz and Buitelaar as well as ours. Examples for *a posteriori* evaluations are the ones of Ciaramita et al., Yamaguchi, but also Schutz and Buitelaar. With respect to the directly comparable approach of Mädche and Staab, our approach gets much higher results in terms of precision or direct matches, i.e. 33.53% compared to 11%. The best *a priori* precision of Schutz and Buitelaar (11.9%) is comparable to the one obtained by Mädche et al. However, the focus of the latter approach was not on learning the right level of generalization. Finally, we would also like to draw the attention to the ASIUM system (Faure and Nedellec, 1998) which addresses the question from a clustering perspective, capturing and generalizing selectional restrictions with respect to hierarchically organized word clusters.

5. Conclusion and Further Work

The contribution of our paper is a systematic analysis of different probabilistic and statistical measures for the purpose of finding the appropriate generalization level for ontological relations extracted from a corpus with respect to a given taxonomy. Our conclusion is that the conditional probability performs better than other measures such as PMI or a χ^2-test. We have so far conducted experiments on the Genia corpus and ontology. In general, we have also observed that it seems quite difficult to find the appropriate generalization due to the fact that the Genia ontology is very small and lacks a reasonable hierarchical structure. Therefore, it remains an open question if our results would transfer to ontologies with a richer structure. The main drawback of our approach is that it is currently restricted to binary relations. Furthermore, the domain and range of a relation can actually not be regarded as independent from each other. However, according to our current observations, an approach to generalizing domain and range dependently could be seriously affected by data sparseness in the Genia corpus. Concerning the approximation of the conditional probabilities, some more elaborate linguistic analysis or even smoothing techniques should be explored. Finally, other structures than verb frames could be considered for deriving relations.

Acknowledgments We acknowledge support from the SmartWeb project[2], funded by the German Ministry of Research and Education.

6. References

S. Abney. 1996. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*, pages 8–15.

M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. 2005. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of IJCAI'05*, pages 659–664.

S. Clark and D.J. Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

D. Faure and C. Nedellec. 1998. 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*, pages 5–12.

U. Hahn and K. Schnattinger. 1998. Towards text knowledge engineering. In *Proceedings of AAAI'98/IAAI'98*, pages 524–531.

M. Kavalec and V. Svátek. 2005. A study on automated relation labelling in ontology learning. In *Ontology Learning from Text: Methods, Evaluation and Applications*, number 123 in Frontiers in Artificial Intelligence and Applications, pages 44–58. IOS Press.

H. Li and N. Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

D. Lin and P. Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

A. Mädche and S. Staab. 2000. Discovering conceptual re-

[2] http://www.smartweb-project.org

lations from text. In *Proceedings of ECAI*, pages 321–325.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Language Processing*. MIT Press.

D. McCarthy. 1997. Word sense disambiguation for acquisition of selectional preferences. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 52–60. Association for Computational Linguistics.

T. Ohta, Y. Tateisi, and J. Kim. 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the 10th Internationl Conference on Human Language Technology (HLT)*.

P. Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*

F. Ribas. 1995. On learning more appropriate selectional restrictions. In *Proceedings of the 7th Conference of the European chapter of the Association for Computational Linguistics (EACL)*, pages 112–118.

A. Schutz and P. Buitelaar. 2005. Relext: A tool for relation extraction from text in ontology extension. In *Proceedings of the International Semantic Web Conference*, pages 593–606.

H. Schütze. 1993. Word space. In *Advances in Neural Information Processing Systems 5*, pages 895–902.

T. Yamaguchi. 2001. Acquring conceptual relationships from domain-specific texts. In *Proceedings of the IJCAI Workshop on Ontology Learning*.