Book Title
Book Editors
IOS Press, 2003

A Study on Automated Relation Labelling in Ontology Learning

Martin KAVALEC and Vojtěch SVÁTEK
Department of Information and Knowledge Engineering
University of Economics, Prague
W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic

Abstract. Ontology learning from texts has been proposed as a technology helping ontology designers in the modelling process. Within ontology learning, the discovery of non-taxonomic relations is understood as the problem least addressed. We propose a technique for extraction of lexical items that may give cue in assigning semantic labels to otherwise 'anonymous' non-taxonomic relations. The technique has been implemented as extension to the existing *Text-to-Onto* tool. Experiments have been carried out on a collection of texts describing tour destinations as well as on a semantically annotated general corpus. The paper also discusses evaluation aspects of relation labelling, among which the distinction of prior and posterior precision looks as most important.

Keywords. Ontology learning, non-taxonomic relation, evaluation.

1. Introduction

Among the three core subtasks of *ontology learning* systematically examined in [12]—lexical entry extraction (also viewed as concept extraction), taxonomy extraction and *non-taxonomic relation extraction*—the last is considered as most difficult. Discovered non-taxonomic relations are *labelled* by a human ontology engineer and become part of an ontology; empirical studies however suggest that ontology engineers may not always easily label a relation between two general concepts, since various relations among instances of the same general concepts are possible [12]. For example, when *some* relation between the concepts 'Company' and 'Product' is detected in textual data, multiple interpretations are at hand: a company may not only *produce* but also *sell*, *consume* or *propagate* a product. The same problem has been witnessed for the medical domain [2]: although a strong relation between the concept 'Chemical/Drug' and 'Disease/Syndrome' was identified in a corpus of medical texts, it was not obvious whether this was mainly due to the semantic relation 'treats', 'causes' or other. Moreover, even if the semantics is clear, it might still be hard to guess which among synonymous labels (e.g. 'produce', 'manufacture', 'make'...) is preferred by the community. *Lexical items* extracted from relevant texts thus may give an important indication for a relevant choice.

There is agreement in the NLP community that relational information is, at sentence level, typically conveyed by *verbs*[1]. The basic idea of this paper is to select verbs (or simple verb phrases) frequently occurring in the context of each concept association. The *concept-concept-verb triples* are then ordered by a numerical measure, and (verbs from) the top ones are candidates for relation labels of the given concept associations. The results of labelling can be evaluated in a similar way as those of other ontology learning tasks, i. e. in terms of precision and recall, though the problems related to construction of a reference model are a bit more severe.

The article is organised as follows. Section 2 explains the principles of our method, suggests quantitative criteria for choosing lexical items as relation labels and describes the implementation of the method. Section 3 outlines a methodology of performance evaluation. Section 4 presents and discusses the results of experiments in the tourism domain. Similarly, section 5 presents and discusses the results of experiments with *SemCor*, a semantically-tagged general corpus. Section 6 reviews related work. Finally, section 7 wraps up the paper and outlines possibilities for future work.

2. Principles and Implementation of Label Extraction

A standard approach to *relation discovery* in text corpora is derived from *association rule learning* [1], originally applied on relational data. In the *text-mining setting* two (or more) lexical items are understood as belonging to a *transaction* if they occur together in a document or other predefined unit of text; frequent transactions are output as *associations* among these items. Furthermore, ontology learning tools discover binary relations not only for lexical items but also for ontological concepts [13]. This presumes existence of a semantic *lexicon* (mapping lexical items to underlying concepts) and preferably a *concept taxonomy*, which enable aggregation of relation instances along the 'is-kind-of' and 'is-a' axes.

Our extended notion of transaction assumes that the 'predicate' of a non-taxonomic relation can be characterised by *verbs* frequently occurring in the neighbourhood of pairs of lexical items corresponding to associated concepts. Information about the verbs is present in the texts, but it gets lost when the texts are transformed to a set of concept co-occurrences.

Definition 1. VCC(n)-transaction holds among a verb v, concept c_1 and concept c_2 iff c_1 and c_2 both occur within n words from an occurrence of v.

Good candidates for labelling a non-taxonomic relation between two concepts are the verbs frequently occurring in VCC(n) transactions with these concepts, for some 'reasonable' n. In the experiments described further we heuristically set n to 8, which takes into account possible articles, prepositions, adjectives or nested clauses in a sentence. With too small n we would lose some important relations between concepts or candidates for labels of such relations. Furthermore, the counts

[1] At the level of noun phrases, this role is dominantly played by prepositions. They however only cover a limited set of domain-neutral relations such as parthood or adjacency.

for estimating co-occurrence probabilities would be too low and the probablity estimates would be unreliable. On the other hand, with too high n, many unrelated items would be considered as related, which would introduce noise. Eight words would probably be too large a distance in general language processing. We however do not count noun-verb (or noun-noun) pairs, but concept-verb (or concept-concept) pairs instead, i.e. only occurrences of terms contained in the ontology lexicon are considered.

A very simple measure of association between a verb and a concept pair is *conditional frequency* (empirical probability)

$$P(c_1 \wedge c_2/v) = \frac{|\{t_i|v, c_1, c_2 \in t_i\}|}{|\{t_i|v \in t_i\}|} \tag{1}$$

where $|.|$ denotes set cardinality, and t_i are the VCC(n)-transactions. It helps to find concept pairs possibly associated with a given verb. However, conditional frequency of a pair of concepts given a verb is not the same as conditional frequency of a *relation* between concepts given a verb. A verb may occur frequently with each of the concepts, and still have nothing to do with any of their mutual relationships. For example, in our first experimental domain, lexical items corresponding to the concept 'city' often occurred together with the verb 'to reach', and the same held for lexical items corresponding to the concept 'island', since both types of location can typically be reached from different directions. Conditional frequency $P(City \wedge Island/'reach')$ was actually higher than for verbs expressing true semantic relations between the concepts, such as 'located' (a city is located on an island). To tackle this problem, we need a measure expressing the *increase* of conditional frequency, as defined in (1), compared to frequency expected under assumption of *independence* of associations of each of the concepts with the verb. Our heuristic 'above expectation' (AE) measure is:

$$AE(c_1 \wedge c_2/v) = \frac{P(c_1 \wedge c_2/v)}{P(c_1/v).P(c_2/v)} \tag{2}$$

(the meaning of $P(c_1/v)$ and $P(c_2/v)$ being obvious). This measure resembles the 'interest' measure (of implication) suggested by Kodratoff [11] as operator for knowledge discovery in text[2]. The 'interest' however merely compares the relative frequency of a pattern (in data) conditioned with another pattern, with its unconditioned relative frequency. Our AE measure, in turn, compares a conditional frequency with the product of two 'simpler' conditional frequencies. We could also reorder the triples by an alternative measure, $AE(v/c_1 \wedge c_2)$: this would yield (possibly even more useful) information on which verbs most typically occur with a certain relation.

By ignoring the *order* of concepts and verb and because of *stemming*, passive and active sentences are treated equally in our approach: this avoids the use of deep parsing. Sentences such as 'Many tourists visit this museum...' and 'This museum is visited by many tourists...' yield both the same triple ('mu-

[2]There is also some similarity with statistical measures such as χ^2. These however involve applicability conditions (regarding the sample size) that are hard to meet in ontology learning, where a high number of relatively infrequent features have to be examined.

seum','tourist', 'visit'), which is desirable. On the other hand, we cannot capture the differences in meaning of sentences like 'A company was hired by a person to accomplish some task' vs. 'A company hired a person to accomplish some task'. It however seems that achieving higher frequencies for concepts and labels is more important, since in the end, a human designer eventually judges the resulting triple. In our example, we expect that the ontology engineer knows that a company may hire a person as well as a person may hire a company, and will appropriately model both relationships in the ontology. Although comprehensive text analysis addressing the aspects of tense might increase usability towards the ontology engineer, the costs would probably not outweigh its benefits.

The computation of VCC(n) transactions and associated frequency measures has been implemented as a new module of the *Text-to-Onto tool* [14]. Resulting concept-concept-verb triples are shown in a separate window popping up from its parent window of 'bare' relation extractor, upon choosing one or more among the relations. In addition, complete results are output to the textual protocol.

3. Performance Evaluation Techniques

A straightforward evaluation technique (see [12] as well as most papers in this volume) is to compare the results of labelling with relation names from a *reference* ('gold standard') ontology created by human evaluators upon reading/browsing a sample of texts. The *precision* and *recall* measures, well-known from information retrieval, can be used to quantify the results. However, as mentioned above, finding suitable names for non-taxonomic relations is more tedious for humans than just listing concepts or even building a concept taxonomy. Moreover, the reliability of 'gold standard' design can be assured, for other tasks, by presenting to the human designer an (almost) exhaustive list of candidate patterns, such as frequent terms (for concept extraction) or concept pairs (for suggestion of taxonomic or anonymous non-taxonomic relations). Names of relations, on the other hand, are linked to lexical items much more loosely than names of concepts[3]: partly because they are not reflected at the lexical level at all, partly because they are dispersed in large synonym sets, and partly because they only pertain to a small subset of occurrences of a term[4]. By consequence, many 'correct' relations would presumably be missing in the reference ontology. An evaluation method exclusively relying on matching relation names from reference ontology with subsequently learnt labels thus might improperly penalise the labelling tool in terms of precision. The solution is to employ two types of precision[5]: *prior* (with respect to reference ontology built prior to learning) and *posterior* (with respect to posterior evaluation of learning results). The latter may be subjectively biased (since the expert may directly control the evaluation result) but makes up for human omissions.

[3]For example, Maedche [12] showed that only 10-15% of human-provided relation labels were found among extracted lexical items, versus 20-25% for concept labels.

[4]While the first two aspects also represent inherent limits for any labelling tool, the third is a specific hindrance to reference ontology design based on a set of extracted frequent terms.

[5]Recall, on the other hand, can only be computed as 'prior'.

Let us now elaborate on this general idea towards a possible *procedural scenario*. Given a previously extracted collection of concepts C, arranged into a taxonomy, the evaluation of extracted relation labels may look as follows:

1. A domain expert *suggests possible named relations* for all pairs of concepts. We thus obtain a set of reference (concept-concept-label) triples that forms, together with the original taxonomy, a *reference ontology*. Since the number of such pairs might be large, only a subset of concepts, $C^* \in C$, could actually be used. Low-level concepts should be pruned as relations among them are less likely to achieve sufficient frequency counts. On the other hand, a few top-level concepts might be pruned as well in some situations since the interpretation of associations among them would be too uncertain. Obviously, the concept-pruning strategy impacts the evaluation results.
2. The labelling tool to be evaluated is run on the document collection and suggests a set of labels for each concept pair from C.
3. The empirical labels are *compared* for equality or synonymy with labels suggested by the expert. The comparison can be carried out either merely by human judgement or by using a lexical resource such as WordNet.
   One of the following types of (non-)match with the reference ontology is identified for any learnt concept-concept-label triple $t = (c_1, c_2, lab)$ such that $c_1, c_2 \in C^*$:
   - t directly matches some *reference triple* (concept pair with verb suggested by expert) $t' = (c_1, c_2, lab')$, i.e. lab and lab' are synonyms or (provided human judgement is used) reflect the same relation between c_1 and c_2
   - t could be matched with a reference triple if c_1 and/or c_2 are properly generalised/specialised in the concept taxonomy
   - t could be matched with a reference triple if lab is replaced with a hyper/hyponym (this would only work if a proper lexical resource is used)
   - combination of the previous cases
   - no match can be found even across taxonomies of both types.

   These situations can be used to compute both *prior precision* and *recall* of labelling, with respect to the set of triples in the reference ontology. Prior precision is the proportion of learnt triples that match some reference triple. Prior recall is the proportion of reference triples that match some learnt triple. Partial match via taxonomies (detected in the previous phase) can either be taken into account or not.
4. Learnt triples $t = (c_1, c_2, lab)$ not (or incompletely) matching with reference triples, i.e. such that either $c_1 \notin C^*$, $c_2 \notin C^*$, or lab is not synonym of lab' from any reference triple $t' = (c_1, c_2, lab')$, are *submitted to the expert* for posterior evaluation.
5. The expert may declare some of the non-matching learnt triples as relevant, and *augment* accordingly the set of correct hits. Two different augmentation variants are possible, 'strict' and 'relaxed':
   - *Strict augmentation*: a triple only becomes relevant if it should have been part of the reference ontology, i.e. the non-match was due to omission.

- *Relaxed augmentation*: a triple always becomes relevant if the expert judges it as a meaningful relation; it may thus not necessarily be relevant for the application domain of the ontology.

6. *Posterior precision* is computed as proportion of reference triples that are marked as correct hits.

Note that the distinction of prior and posterior precision can in principle be applied on any ontology learning task; in relation labelling, however, the span between the two is potentially widest due to (often) numerous alternative relations between the same concepts.

In the experiments described in sections 4 and 5, we only applied fragments of the above scenario, mainly due to small size and specific nature of data.

4. Experiment in Tourism Domain: Lonely Planet Collection

4.1. Problem Setting

For the first experiment we adopted the Lonely Planet text collection[6]: 1800 short documents in English, about 5 MB overall[7]. These are free-text descriptions of various tourist destinations encompassing geography, history and available leisure activities. Our goal was to verify to what extent such a text collection can be used as support for discovering and *labelling* non-taxonomic relations for an ontology of the domain. Such an ontology could be used for diverse purposes, from ad-hoc question answering about world geography to tour recommendation applications.

Non-taxonomic relation extraction is a task typically superimposed over several other tasks, which can be carried out via manual modelling or inductively from text: lexical entry extraction, mapping of lexical entries to concepts, and taxonomy building:

- In *Text-to-Onto*, *lexical entry extraction* has previously been used for discovery of potential *concept* labels, based on the well-known TFIDF (term frequency - inverse document frequency) measure. In contrast, our goal was *relation* labelling, which is also a form of lexical entry extraction but requires a more focused approach. Since our hypothesis was that 'relational' information is most often conveyed by verbs, we integrated a *part-of-speech* (POS) tagger into the process of frequent transaction discovery[8]. About 75000 verb occurrences were identified in the collection.
- Although mapping *lexical items to concepts* can be accomplished automatically (via information extraction) in principle, the reliability of man-made resources is significantly higher. We thus adopted portions of the *TAP knowledge base*[9] recently developed at Stanford. TAP is a large repository

[6] `http://www.lonelyplanet.com/destinations/`

[7] The same dataset was later used in other experiments with the *Text-to-Onto* tool [6]

[8] The same POS tagger, QTag `http://www.english.bham.ac.uk/staff/omason/software/qtag.html`, was previously used in *Text-to-Onto* for term extraction but not in the context of relation discovery.

[9] `http://tap.stanford.edu`

of lexical entries, such as proper names of places, companies, people, but also names of sports, art styles and other less traditional 'named entities'. It has previously been used for automated annotation of web pages [7] but its use as a lexicon for ontology learning is novel.

- TAP includes a simple *taxonomy*, which is however not compatible with standard upper-level ontologies and contains ontologically unsound constructs. We therefore (manually) combined the TAP taxonomy with a small hand-made tourism ontology, and slightly extended it via the *Text-to-Onto* term extraction facility. Although *Text-to-Onto* also contains an automatic taxonomy-building tool, we did not use it to prevent error chaining from one ontology learning task to another.

4.2. Analysis and Results

The whole analysis consisted of several phases, in which we used different components of *Text-to-Onto*. The output of earlier phases was stored and subsequently used for multiple (incl. debugging) runs of the last phase.

1. First, occurrences of ontology concepts (i.e. lexicon entries) were found in text and stored in an index. For all 157 concepts, there were about 9300 such entries with about 70000 occurrences.
2. Next, we used the POS tagger to identify the occurrences of verb forms in the text. About 75000 verb occurrences were identified; they were stored in another index.
3. Finally, we compared the indices from step 1 and 2, recorded the VCC(n)-transactions, and aggregated them by triples.

Table 1 lists the 24 concept-concept-verb triples with $AE(c_1 \wedge c_2/v)$ higher than 100% (ordered by this value); triples with occurrence lower than 3, for which the relative frequencies do not make much sense, have been eliminated. The symbol $C(v, c_1, c_2)$ stands for $|\{t_i|v, c_1, c_2 \in t_i\}|$, i.e. how many times the verb occurred close enough to both concepts.

4.3. Evaluation

We can see that triples with high $AE(c_1 \wedge c_2/v)$ (even those with low absolute frequencies, 4 or 5) correspond to meaningful semantic relations, mostly topo-mereological ones: an island or a country is located in a world-geographical region (wg_region), a country 'is a country' of a particular continent and may be located on an island or consist of several islands. On this small result set, we can simulate the evaluation strategy outlined in section 3. For simplicity (and to minimise subjective bias), we only chose as *reference ontology* the set of obvious topo-mereological relations among geographical concepts. For the most frequent six concepts of this kind (City, US City, Country, Island, Continent, World Geographic Region), we identified 17 concept-concept-relation triples that are likely to frequently occur in reality: 14 topological ones (i.e. an object is located within another object, under transitive closure) and 3 mereological ones (i.e. an object consists of other objects). The reference ontology is at Fig. 1; line arrows stand

Table 1. Concepts with suggested labels for *Lonely Planet* collection

c_1	c_2	v	$C(v, c_1, c_2)$	$P(c_1 \wedge c_2/v)$	$AE(c_1 \wedge c_2/v)$
island	wg_region	*locate*	3	0.95%	750.00%
country	wg_region	*locate*	10	3.17%	744.68%
continent	country	*is_country*	26	10.12%	431.10%
us_city	wg_region	*locate*	4	1.27%	350.00%
country	island	*made*	5	1.68%	270.42%
country	island	*locate*	5	1.59%	239.36%
country	island	*consist*	10	7.41%	234.78%
museum	us_city	is_home	3	1.74%	234.55%
country	island	*comprise*	6	5.56%	200.62%
country	tourist	enter	6	2.79%	176.95%
country	island	*divide*	5	3.88%	172.46%
island	us_city	*locate*	3	0.95%	168.75%
city	stadium	known	9	1.25%	165.69%
city	country	allow	24	13.71%	152.89%
city	tourist	is_city	9	1.74%	151.61%
country	us_city	*locate*	9	2.86%	150.80%
city	country	*is_settlement*	6	16.22%	148.00%
island	us_city	connect	3	2.86%	140.00%
country	island	populate	5	6.02%	139.73%
city	island	*locate*	8	2.54%	131.39%
city	country	reflect	5	8.06%	117.42%
city	country	grant	4	12.90%	105.98%
city	park	is_city	11	2.13%	104.23%
city	country	stand	8	5.06%	104.03%

for 'located in', full arrow for 'is-a' and diamond arrows for 'consists of'. We could then compute (prior) *precision*, *recall*, and, finally, *F-measure* (harmonic mean of precision and recall [17]) with respect to the reference ontology. For simplicity, we did not take concept taxonomy (in this case, a single is-a link) nor verb hypero/hyponymy into account; only verbs that directly reflect the given relation (italicised in Table 1) were considered. Furthermore, there are relations that are not included in the reference ontology but still make sense, for example the 'entering' relation between the concepts of Tourist and Country. If we choose the relaxed variant of *augmentation*, we keep such cases as correct hits rather than as misses. We can then compute the *posterior precision*. Fig. 2 shows the recall and both types of precision in a single graph, while Fig. 3 shows the F-measure (the X-axis always corresponds to increasing number of triples in the descending order of AE measure). The F-measure value sharply increases as long as the values of AE measure are in the order of multiple hundreds, then less sharply for values around 130-230%, and finally monotonically decreases when approaching to 100% (i.e. 'equal-to-expectation' value). The sample size was however so small that no general conclusions could be drawn from these figures.

Set aside the solid recall on topo-mereological relation labels, the total number of labels extracted from the 5MB corpus was definitely not impressive. This can

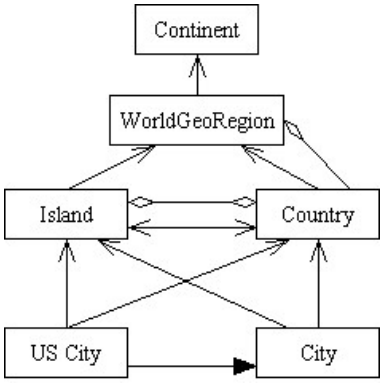

Figure 1. Reference ontology for Lonely Planet experiment

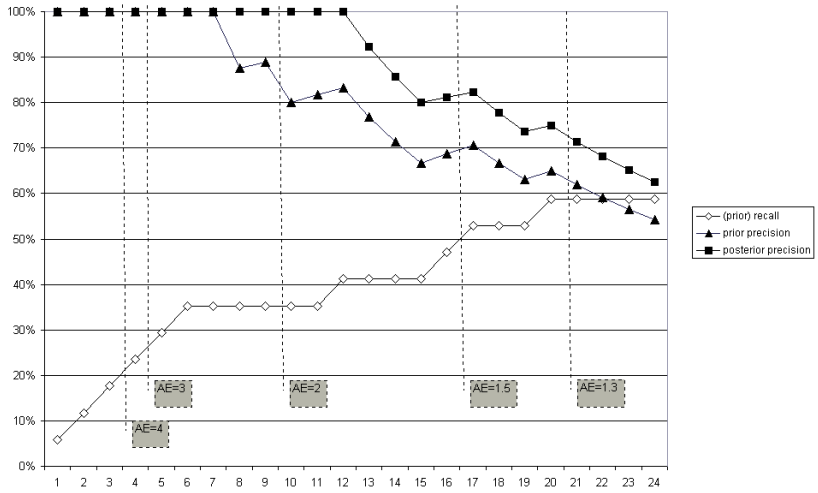

Figure 2. Recall and (prior and posterior) precision in Lonely Planet experiment

be partially attributed to the following:

- *Sparseness of concept taxonomy.* The TAP-based taxonomy was not a true ontology of the domain, and was rather sparse.
- *Sparseness of lexicon.* The lexicon only covered a part of the relevant lexical space. It listed many names of places (often only appearing in a single document) but few names of activities for tourists or art objects (reusable across many documents). Better coverage would require either comprehensive lexicons (some can also be found on the web) or heavy-weighted linguistic techniques such as anaphora resolution, since the geographical entities initially introduced in the text are often referred to by pronouns.
- *Semantic ambiguity of terms.* Ambiguous words were assigned all possible meanings, which of course added *noise* to the data.

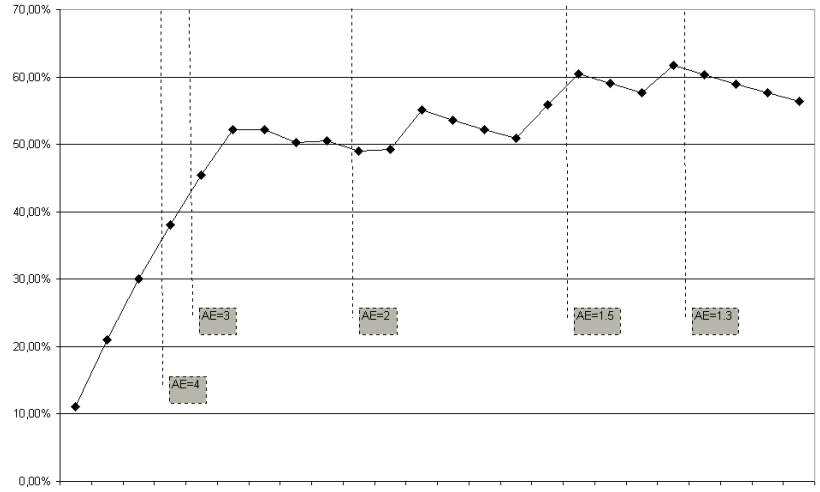

Figure 3. (Prior) F-measure in Lonely Planet experiment

- *Style of underlying text.* The Lonely Planet documents are written in a free style: the same relation is often expressed by different verbs, which decreases the chance of detecting the most characteristic one.
- *Performance of POS tagger.* Sometimes, the tagger does not properly categorise a lexical item. For example, a verb associated with concept *Country* was 'cross'; some of its alleged occurrences however seemed to be adverbs (e.g. 'cross-country skiing' or 'Cross-country Touring Center').
- *Performance of concept extractor.* Since relation extraction was superimposed over (automated) concept extraction, results of the former were negatively influenced by the flaws of the latter.

More detailed results of the *Lonely Planet* experiment can be found in [10].

5. Experiments with Semantically Tagged Corpus

5.1. Problem Setting

In order to overcome some difficulties arisen in the previous experiment, we adopted *SemCor*[10]: a part of the Brown corpus[11], semantically tagged with WordNet[12] senses. All open word classes (nouns, verbs, adjectives and adverbs) are mapped to their WordNet senses. Advantages over an ad hoc document collection such as Lonely Planet immediately follow from reduced ambiguity:

[10] `http://www.cs.unt.edu/~rada/downloads.html`
[11] `http://helmer.aksis.uib.no/icame/brown/bcm.html`
[12] `http://www.cogsci.princeton.edu/~wn`

1. We can use the WordNet hierarchy to lift the tagged terms to *concepts* at an arbitrary level of abstraction. There is thus no need for automatic (and error-prone) frequency-based concept extraction.
2. Similarly, we can aggregate the *verbs* along the hierarchy and thus overcome their sparseness of data.
3. We can evaluate our approach without the impact of a POS tagger, which also exhibited a significant error rate in the previous experiment.

Since *SemCor* is a small corpus with very broad scope, we confined ourselves to three very general concepts to avoid data sparseness: *Person*, *Group* and *Location*[13]. We identified each of them with the WordNet synset containing the word sense person#1 (or group#1 or location#1, respectively) and all its hyponyms. Any word tagged with a WordNet sense that could be generalised to the synset containing person#1 was thus considered as occurrence of Person (and the like). This way we found 14613 occurrences for Person, 6727 for Group and 4889 for Location. The corpus contains 47701 sense-tagged verb occurrences. In all three experiments below, we set the minimal absolute frequency of triples to 5, to filter out the cases where the relative frequencies were skewed because of sparse data.

5.2. Analysis and its Results

In the first experiment with *SemCor* we grouped the verbs directly by the *synset* they belong to (i.e. all occurrences of verbs from one synset counted together); this yielded 4894 synsets. Table 2 shows the top synsets according to the AE score (only considering those with $AE \geq 2.5$), for the Person-Group pair. In the second experiment we generalised each verb by taking its (first-level) *hypernym synset*; we obtained 1767 synsets. Top ones (again considering those with $AE \geq 2.5$) for the Person-Group pair are in Table 3. In the third experiment we attempted to introduce some 'domain bias' by separately processing two *sub-collections* of *SemCor*, news articles and scientific texts, each representing about 15% of the original corpus. We generally observed dissimilar distributions of verb synsets (e.g. news articles concerned 'moving', 'communicating', 'leading', while scientific texts rather dealt with 'observing', 'proposing' or 'transforming') however, only a fraction of verbs suggested as labels for a particular relation was relevant. This was obviously due to data sparseness, even in the hypernym synset setting.

5.3. Evaluation

Since building a 'reference ontology' corresponding to the coverage of a generic corpus is unconceivable, we cannot evaluate the labels by means of prior precision and recall. The only remaining measure is then *posterior precision* based on subjective evaluation (i.e. 'relaxed augmentation of empty reference ontology'). We considered as positive hits all cases where *at least one member* of the verb synset corresponded to a meaningful relation among the concepts that would be worth

[13] Admittedly, the combination of a generic corpus and a three-class target 'ontology' does not approximate real-world (say, business) ontology learning settings very well. It was only meant for 'in vitro' evaluation of the method.

Table 2. Suggested relations between Person and Group – verb synset version

Verb synset	$C(v, c_1, c_2)$	$AE(c_1 \wedge c_2/v)$
head, lead	10	4.43
act as	13	4.36
leave, depart, pull up stakes	7	4.08
decrease, diminish, lessen, fall	6	3.54
submit, state, put forward, posit	9	3.44
serve	11	3.44
form, organize, organise	10	3.41
stage, present, represent	6	3.22
collaborate, join forces, cooperate, get together	8	2.95
include	25	2.68
meet, ran into, encounter, run across, come across, see	10	2.68
meet, gather, assemble, forgather, foregather	5	2.59

Table 3. Suggested relations between Person and Group – verb hypernym version

Verb synset	$C(v, c_1, c_2)$	$AE(c_1 \wedge c_2/v)$
serve, function	13	4.36
attack, assail	6	3.53
meet, ran into, encounter, run across, come across, see	10	2.74
be, follow	11	2.58

modelling in some domain ontology. To assess the impact of verb abstraction, we separately measured the precision for original and abstracted synsets. We only list the graphs for Person-Group pair, in Fig. 4, for labels ordered in the decreasing order of AE measure. It seems that the precision is again decreasing more steeply for triples with AE measure under approx. 130%, although some improper labels cause an abrupt decrease near the beginning. Interestingly, most of such highly-scored false hits are related to *communication* (such as 'state', 'write', 'publish', 'announce', 'remark'). We can hypothesise that especially in news articles, such verbs typically occur near statements involving both persons and groups, yet have nothing to do with the relationship *among* persons and groups. The *hypernym version* had better precision. The most likely reason might be that most 'communication' verbs mentioned above have broad hypernyms such as 'create', which can be considered as proper labels for the Person-Group pair. The labels for other two pairs (Person-Location, Group-Location) had lower precision. While some triples were relevant (such as "Person - born - Location" or "Group - reach - Location"), many other only seemed to reflect the fact that *events* involving persons and/or groups are often said to happen in a particular location. The number of 'correct hits' was hence too low to evaluate the trend of the precision curve.

6. Related Work

Many approaches to relation learning from text do not distinguish much between *relations* and *relation instances*, in the set-theoretic sense. Lexical labels are of-

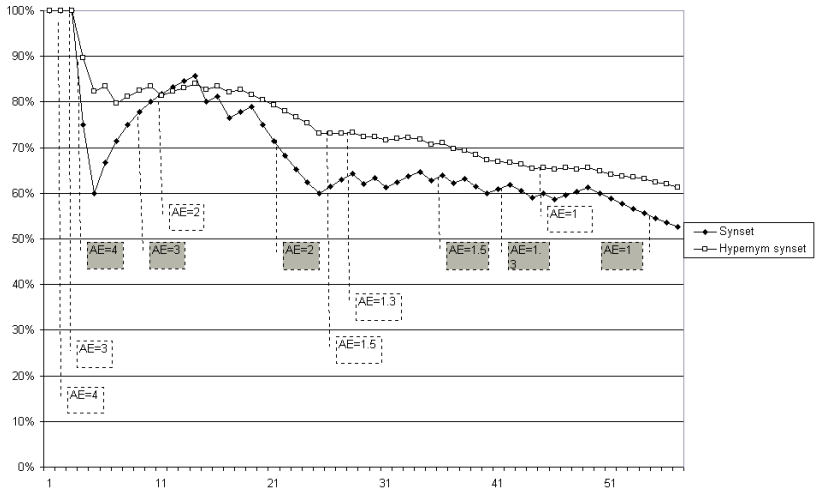

Figure 4. Posterior precision in SemCor experiment

ten directly assigned to statements about concrete pairs of entities, i.e. relation instances. Instances are however usually not expected to be part of an ontology. This research should be viewed as ontology *population*, rather than learning. In contrast, we focus on relations that *possibly* hold among (various instances of) certain ontology concepts. The design of relations is a creative task: it *should* be accomplished by a human, for whom we only want to offer partial support. Yet, many partial techniques are similar. Finkelstein&Morin [9] combine 'supervised' and 'unsupervised' extraction of relationships between terms; the latter (with unspecified underlying relations) relies on 'default' labels, under assumption that e.g. the relation between a Company and a Product is always 'produce'. They also mention the possibility to use specific words (such as verbs) involved in the relationship but do not elaborate it further. Byrd&Ravin [5] assign the label to a relation (instance) via specially-built finite state automata operating over sentence patterns. Some automata yield a pre-defined relation (e.g. *location* relation for the '-based' construction) while other pick up a promising word directly from the analysed sentence. Labelling of proper relations is however not addressed, and even the 'concepts' are a mixture of proper concepts and instances. The *Adaptiva* system [3] asks the user to choose a relation from the ontology and then interactively learns its recognition patterns. Although the goal is to *recognise* relation instances in text, the interaction with the user may also give rise to new proper relations. Such frequent interaction however does not pay off if the goal is merely to *find labels* for important domain-specific relations to which the texts refer, as in our case. The *Asium* system [8] synergistically builds two hierarchies: that of concepts and that of verb sub-categorisation frames (an implicit 'relation taxonomy'), based on co-occurrence in text . Verbs co-occurring with concepts in text are used to cluster the concepts, and vice versa. There is however no direct support for conceptual 'leap' from a 'bag of verbs' to a named relation, which we have thanks to integration of our technique into the whole *Text-to-Onto* environment.

Another line of work, more firmly grounded in ontology engineering, systematically seeks new *unnamed* relations in text. Co-occurrence analysis with limited attention to sentence structure is used, and the results filtered via frequency measures as in our approach. As mentioned before, in prior work on *Text-to-Onto* [13], the labelling problem was left to the ontology engineer. The same holds about the non-taxonomic relation component of *DODDLE* [16], which only differs by a more sophisticated way of transaction construction. In the *OntoLearn* project [15], WordNet and FrameNet mappings are used to automatically assign relations from a predefined set (such as 'similar' or 'instrument').

Interesting is the *OntoLT* plug-in to Protégé [4], which does not distinguish ontology learning tasks such as creation of classes, slots or instances at the architectural level but rather as action parts of user-definable rules. Its input is a corpus that is linguistically annotated by means of another automatic tool (parser): it thus does not rely on surface patterns. The words are filtered for domain specificity (using the χ^2 measure) in the pre-processing phase. Ontology learning corresponds to slot creation; the lexical label for a new slot is directly transferred from (even a single occurrence of) the linguistic predicate for the phrase on which a slot-creation rule is applied.

7. Conclusions and Future Work

Our experiments suggest that ontology learning from text may be used not only for discovering ('anonymous') relations between pairs of concepts, but also for providing potential lexical *labels* for these relations. Verbs, merely identified by POS tagging (i.e. without structural analysis of the sentence) can be viewed as first, rough, approximation of such labels. Serious problems however are the sparseness of data (due to multiple reasons) and domain-dependency of the labels. The experiment with semantically-tagged corpus suggests that referring to the *right word sense* improves the quality of relation labelling, and using *more abstract verbs* (by their generalization via WordNet) may help too, when applied carefully.

The quality of results on the SemCor corpus was comparable to the Lonely Planet experiment despite the smaller and broader corpus; we assume that the presence of semantic information (word senses) made up for the smaller size of the corpus. Although we usually lack word sense information in real-world settings, it is often possible to restrict the senses of words with respect to a narrow domain, for which we build the ontology. In particular, polysemous verbs typically become monosemous in the context of domain-specific applications. In addition, existing methods for disambiguation of named entities could be applied in some cases.

A problematic point of the method is the *direct mapping* from co-occurrences of terms onto 'deep' ontological relations. In particular the SemCor experiment indicated that the method improperly suggests verbs that typically occur in some larger semantic context involving (among others) the two concepts in question but do not correspond to an immediate relation between them. In the future, we plan to make the method more *linguistic-aware*, i.e., to employ a deep or shallow or parser to determine the (syntactically) most appropriate verb within the transaction. We would like to determine whether the overhead of shallow

parsing will be outweighed by better precision. The most important task for the future is however to apply our method to a *domain-specific* collection of texts relevant to a clearly-defined application.

Acknowledgments

The research is partially supported by grant no.201/03/1318 of the Czech Science Foundation. Its initial part was carried out during M. Kavalec's stay at FZI Karlsruhe; we owe our thanks to Alex Maedche, who took part in this phase.

References

[1] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. SIGMOD-93, 207–216
[2] Bodenreider, O.: Medical Ontology Research. A report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications, National Library of Medicine 2001.
[3] Brewster, C., Ciravegna, F., Wilks, Y.: User-Centred Ontology Learning for Knowledge Management In: 7th Int'l Conf. Applications of Natural Language to Information Systems, Stockholm, LNAI, Springer 2002.
[4] Buitelaar, P., Olejnik, D., Sintek, M.: A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. In: Proc. ESWS-04, Heraklion 2004.
[5] Byrd, R., Ravin, Y.: Identifying and Extracting Relations in Text. In: Proceedings of NLDB 99, Klagenfurt, Austria, 1999.
[6] Cimiano, P., Automatic acquisition of taxonomies from text: FCA meets NLP. In: Proceedings of the International Workshop on Adaptive Text Extraction and Mining (ATEM03), Cavtat 2003.
[7] Dill, S. et al.: SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In: Proc. WWW2003, Budapest 2003.
[8] Faure, D., Nédellec, C.: ASIUM: Learning subcategorization frames and restrictions of selection. In: ECML'98, Workshop on Text Mining, 1998.
[9] Finkelstein-Landau, M., Morin, E.: Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods. In: Int'l Workshop on Ontological Engineering on the Global Information Infrastructure, Dagstuhl 1999.
[10] Kavalec, M., Maedche, A., Svátek, V.: Discovery of Lexical Entries for Non-Taxonomic Relations in Ontology Learning, In: SOFSEM'04, Springer LNCS, 2004.
[11] Kodratoff, Y.: Comparing Machine Learning and Knowledge Discovery in DataBases: An Application to Knowledge Discovery in Texts. In: ECCAI summer course, Crete July 1999, LNAI, Springer 2000.
[12] Maedche, A.: Ontology Learning for the Semantic Web. Kluwer, 2002.
[13] Maedche, A., Staab, S.: Mining Ontologies from Text. In: Proc. EKAW'2000, 2000.
[14] Maedche, A., Volz, R.: The Text-To-Onto Ontology Extraction and Maintenance System. In: ICDM-Workshop on Integrating Data Mining and Knowledge Management, San Jose, California, USA, 2001.
[15] Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F.: Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System. In this volume.
[16] Sugiura, N., Shigeta, Y., Fukuta, N., Izumi, N., Yamaguchi, T.: Towards On-the-Fly Ontology Construction – Focusing on Ontology Quality Improvement. In: 1st European Semantic Web Symposium (ESWS-04), Heraklion, Greece, 2004.
[17] van Rijsbergen, C. J.: Information Retrieval. London, Butterworths, 1979.