Discovery of Lexical Entries for Non–Taxonomic Relations in Ontology Learning

Martin Kavalec[1], Alexander Maedche[2], and Vojtěch Svátek[1]

[1] Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
{kavalec,svatek}@vse.cz
[2] Robert Bosch GmbH, Germany
Alexander.Maedche@de.bosch.com

Abstract. Ontology learning from texts has recently been proposed as a new technology helping ontology designers in the modelling process. Discovery of non–taxonomic relations is understood as the least tackled problem therein. We propose a technique for extraction of lexical entries that may give cue in assigning semantic labels to otherwise 'anonymous' relations. The technique has been implemented as extension to the existing *Text-to-Onto* tool, and tested on a collection of texts describing worldwide geographic locations from a tour–planning viewpoint.

1 Introduction

Ontologies are the backbone of the prospective semantic web as well as of a growing number of knowledge management systems. The difficulty of their manual development is however a significant drawback. Recently, *ontology learning* (OL) from text has been suggested as promising technology for building lightweighted ontologies with limited effort. It relies on combination of shallow text analysis, data mining and knowledge modelling. In [9], three core subtasks of OL have systematically been examined: lexical entry extraction (also used for concept extraction), taxonomy extraction, and *non–taxonomic relation*[3] *extraction* (NTRE), considered as most difficult. The NTRE technique [10] embedded in the *Text–to–Onto* tool [11] of the KAON system[4] produces, based on a corpus of documents, an ordered set of binary relations between concepts. The relations are *labelled* by a human designer and become part of an ontology. Empirical studies [9] however suggest that designers may not always appropriately label a relation between two general concepts (e.g. 'Company' and 'Product'). First, various relations among instances of the same general concepts are possible; for example, a company may not only *produce* but also *sell*, *consume* or *propagate* a product. Second, it is often hard to guess which among synonymous labels

[3] Although it might be useful to distinguish the terms 'relation' and 'relationship' (set of tuples vs. high–level association between concepts), we mostly speak about 'relations' since this term is systematically used in the ontology engineering community.

[4] Karlsruhe Ontology infrastructure, http://kaon.semanticweb.org.

(e.g. 'produce', 'manufacture', 'make'...) is preferred by the community. *Lexical entries* picked up from domain–specific texts thus may give an important cue.

The paper is organised as follows. Section 2 describes the principles of our method. Section 3 presents and discusses the results of an experiment in the tour–planning domain. Section 4 compares our approach with related research. Finally, section 5 summarises the paper and outlines directions for future work.

2 Seeking Labels for Relations in *Text–to–Onto*

2.1 Method Description

The standard approach to *relation discovery* in text corpus is derived from *association rule learning* [1]. 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 their items. *Text–to–Onto*, however, discovers binary relations not only for lexical items but also for ontological concepts [10]. This presumes existence of a *lexicon* (mapping lexical entries to underlying concepts) and preferably a *concept taxonomy*.

Modification of the method, which is the subject of this paper, relies on an extended notion of transaction. Following up with our prior work on lexical entry extraction from business websites [7], we hypothesised that the 'predicate' of a non–taxonomic relation can be characterised by *verbs* frequently occurring in the neighbourhood of pairs of lexical entries corresponding to associated concepts.

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. Very simple measure of association between a verb and a concept pair are conditional frequencies (empirical probabilities)

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

$$P(v/c_1 \wedge c_2) = \frac{|\{t_i|v, c_1, c_2 \in t_i\}|}{|\{t_i|c_1, c_2 \in t_i\}|} \tag{2}$$

where $|.|$ denotes set cardinality, and t_i are the VCC(n)–transactions. The first one helps to find concept pairs possibly associated with a given verb; the second one helps to find verbs possibly associated with a given concept pair.

However, conditional frequency of a pair of concepts given a verb (or vice versa) is not the same as conditional frequency of a *relation* between concepts given a verb (or vice versa). 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 experimental domain, lexical entries corresponding to the concept 'city' often occur together with the verb 'to reach', and the same holds for lexical entries corresponding to the concept 'island', since both types of location can typically be reached from different directions. Therefore, conditional frequencies $P(City \wedge Island/'reach')$ and $P('reach'/City \wedge Island)$ will be relatively high, and might even dominate those of verbs expressing a true semantic relation 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) and (2), compared to frequency expected under assumption of *independence* of associations of each of the concepts with the verb. Our heuristic 'above expectation' (AE) measure thus is, respectively:

$$AE(c_1 \wedge c_2/v) = \frac{P(c_1 \wedge c_2/v)}{P(c_1/v).P(c_2/v)} \tag{3}$$

$$AE(v/c_1 \wedge c_2) = \frac{P(v/c_1 \wedge c_2)}{P(v/c_1).P(v/c_2)} \tag{4}$$

(the meaning of $P(c_1/v)$ etc. being obvious). This measure resembles the 'interest' measure (of implication) suggested by Kodratoff [8] as operator for knowledge discovery in text[5]. The 'interest' however merely compares the relative frequency of a pattern (in data) conditioned with another pattern, with its unconditioned relative frequency. Our AE measures, in turn, compare a conditional frequency with the product of two 'simpler' conditional frequencies.

2.2 Implementation

The computation of VCC(n) transactions and associated frequency measures has been implemented as a new module of *Text–to–Onto tool*. 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. A screenshot of KAON environment is at Fig. 1; note the list of verbs potentially associated with relations between 'Country' and 'City', in the front window. In addition, complete results are output into a textual protocol.

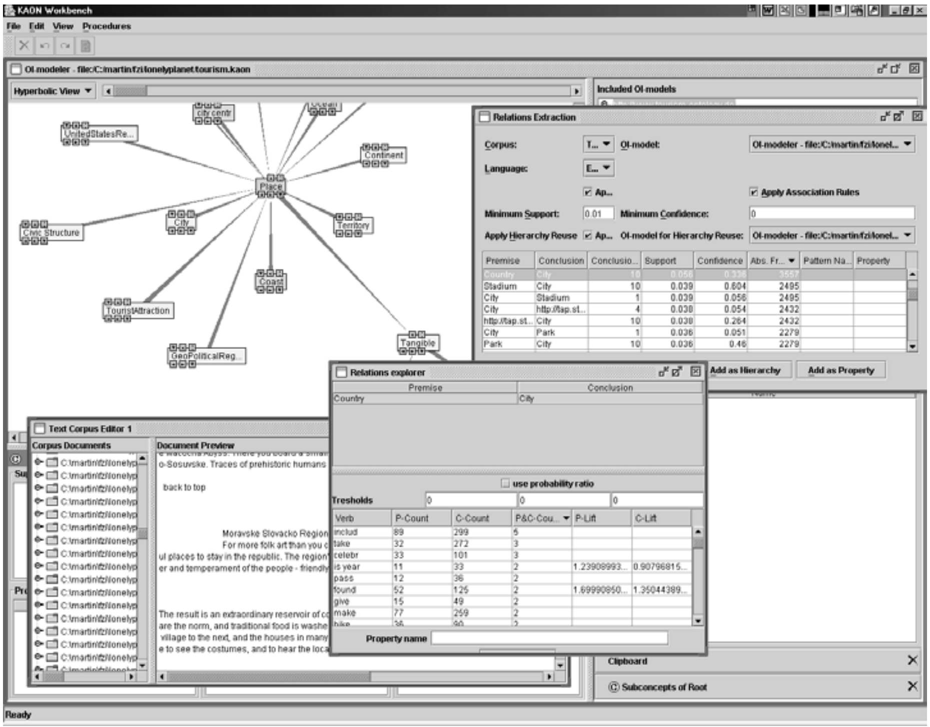
3 Experiments

3.1 Problem Setting

For experiments, we selected the popular domain of tourism. Our text corpus contained web pages from the Lonely Planet website[6]: 1800 short documents in English, about 5 MB overall. These are free–text descriptions of various world locations encompassing geography, history and available leisure activities; there is no systematic information about hotel infrastructure. Our goal was to verify

[5] There is also some similarity with statistical measures such as χ^2. These however involve applicability conditions that are hard to meet in OL, where a high number of relatively infrequent features have to be examined.

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



Corpus: T... OI-model: OI-modeler - file:/C:/martin/fzi/lonel...

Language: E...

☑ Ap... ☑ Apply Association Rules

Minimum Support: 0.01 Minimum Confidence: 0

Apply Hierarchy Reuse ☑ Ap... OI-model for Hierarchy Reuse: OI-modeler - file:/C:/martin/fzi/lonel...

Premise	Conclusion	Conclusio...	Support	Confidence	Abs. Fr...	Pattern Na...	Property
Country	City	10	0.056	0.336	3557		
Stadium	City	10	0.039	0.604	2495		
City	Stadium	1	0.039	0.056	2495		
City	http://tap.st...	4	0.038	0.054	2432		
http://tap.st...	City	10	0.038	0.264	2432		
City	Park	1	0.036	0.051	2279		
Park	City	10	0.036	0.46	2279		

Premise	Conclusion
Country	City

□ use probability ratio

Tresholds 0 0 0

Verb	P-Count	C-Count	P&C-Cou...	P-Lift	C-Lift
includ	89	299	5		
take	32	272	3		
celebr	33	101	3		
is year	11	33	2	1.23908993...	0.90796815...
pass	12	36	2		
found	52	125	2	1.69990850...	1.35044389...
give	15	49	2		
make	77	259	2		
hike	36	90	2		

Property name

Fig. 1. KAON environment with interface for non–taxonomic relation discovery

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 serious tour recommendation applications.

NTRE is a task typically superposed over several other tasks, which can be carried out via manual modelling or OL: 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 (information retrieval) measure, in the whole document collection. 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 embedded a *part–of–speech* (POS) tagger into the process of frequent transaction discovery.
- Mapping *lexical entries to concepts* can hardly be accomplished automatically. We adopted portions of the *TAP knowledge base*[7] recently developed

[7] http://tap.stanford.edu

at Stanford: a large repository of lexical entries—proper names of places, companies, people and the like. It has previously been used for automated annotation of web pages [4] but its use as 'lexicon for OL' was novel.

- TAP includes a simple *taxonomy*, which is however not compatible with standard upper–level ontologies and contains ontologically unsound constructs. We therefore combined the TAP taxonomy with our small hand–made tourism ontology, and slightly 'tweaked' it where needed. Although *Text–to–Onto* also contains an automatic taxonomy–building tool, we did not use it to prevent error chaining from one OL task to another.

3.2 Analysis and its 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, locations of ontology concepts (i.e. lexicon entries) were found in text and stored in an index. There were about 9300 such entries.
2. Next, we used the POS tagger to identify the locations of verb forms in the text; they were stored in another index.
3. We post–processed the POS tags to couple verbs such as 'to be' or 'to have' with their presumed syntactical objects, to obtain more usable verb constructs (these were subsequently handled in the same way as generic verbs).
4. Finally, we compared the indices from step 1 and 2, recorded the VCC(n)–transactions for $n = 8$, and aggregated them by triples. This last phase took about 45 seconds on a 1.8GHz Athlon XP computer.

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[8].

We can see that roughly the first half of triples (even those with low absolute frequencies, 4 or 5) corresponds 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[9], a city may be home of a famous museum etc. However, with $AE(c_1 \wedge c_2/v)$ dropping to about 150 %, the verbs cease to pertain to a relation. This leads us to the heuristics that triples below this value should probably not be presented to the ontology designer.

Note that the table suggests which pairs of concepts should certain verbs be assigned to, as lexical entries for non–taxonomic relations. We could also reorder the triples by an alternative measure, $AE(v/c_1 \wedge c_2)$: this would yield (also quite useful) information on which verbs most typically occur with a certain relation.

[8] Since some of required filtering options were not yet available through the window environment at the moment of performing the experiments, the results were partly obtained via offline analysis of textual protocol produced by *Text–to–Onto*.

[9] Example of *multiple relations* between the same concepts, cf. end of section 1.

c_1	c_2	v	$\lvert\{t_i \mid v, c_1, c_2 \in t_i\}\rvert$	$P(c_1 \wedge c_2/v)$	$AE(c_1 \wedge c_2/v)$
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%

Table 1. Final results of label extraction

The results do not seem too impressive given the amount of underlying material. This however reflect many circumstances independent of the method itself:

- *Richness and relevance of concept taxonomy.* The TAP–based taxonomy was not a true ontology of the domain, and was rather sparse. Construction of a good taxonomy is a demanding task; by complex study in [9], however, it is not as big a challenge as the invention of plausible non–taxonomic relations.
- *Richness and relevance of lexicon.* The lexicon only covered a part of the relevant lexical space. It listed many names of places (most however only appeared 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.
- *Style of underlying text.* The Lonely Planet documents are written in a quite free, expressive style. The same relation is often expressed by different verbs, which decreases the chance of detecting a single, most characteristic one.
- *Performance of POS tagger.* Sometimes, the tagger does not properly categorise a lexical entry. For example, a verb associated with concept *Country* was 'cross'; some of its alleged occurrences however seemed to be adverbs.

4 Related Work

Our work differs from existing research on 'relation discovery' in a subtle but important aspect: in other projects, the notion of 'relation' is typically used for relation *instances*, i.e. statements about concrete pairs of entities: labels are directly assigned to such pairs. Rather than OL in the proper sense (since instances are usually not expected to be part of an ontology), this research should be viewed as *information extraction* (IE). In contrast, we focus on *proper relations*, which *possibly* hold among (various instances of) certain ontology concepts. The number of relations is much lower than the number of their instances but their design is a demanding, creative task. For these reasons, it *can* and *should* be accomplished by a human, for whom we only want to offer partial support.

Yet, many partial techniques are similar. Finkelstein&Morin [6] 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'. Byrd&Ravin [3] 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 from the sentence itself. Labelling of proper relations is however not addressed, and even the 'concepts' are a mixture of proper concepts and instances. The *Adaptiva* system [2], allows the user to choose a relation from the ontology and 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 massive interaction however does not pay off if the goal is merely to *find* important domain–specific relations to which the texts refer, as in our case. The *Asium* system [5] synergistically builds two hierarchies: that of concepts and that of verb subcategorisation frames (an implicit 'relation taxonomy'), based on co–occurrence in text . There is however no direct support for conceptual 'leap' from a 'bag of verbs' to a named relation.

Another stream, more firmly grounded in ontology engineering, systematically seeks new *unnamed* relations in text. Co–occurrence analysis (with little attention to sentence structure) is used, and the results filtered via frequency measures, as in our approach. In prior work on the *Text–to–Onto* project [10], the labelling problem was left upon the ontology designer. In the *OntoLearn* project [12], WordNet mapping was used to automatically assign relations from a small predefined set (such as 'similar' or 'instrument').

5 Conclusions and Future Work

Our experiment suggests that ontology learning from text may be used not only for discovering ('anonymous') relations between pairs of concepts, but also for providing lexical entries as potential *labels* for these relations. Verbs, identified merely by POS tagger (i.e. without structural analysis of the sentence) can be viewed as first, rough, approximation of the desired category of such entries.

Most imminent future work concerns the possibility to immediately verify the semantics of discovered concept–concept–verb triples, via return to the original text. Sometimes the ontology designer might wonder (e.g. assuming a 'borderline' AE measure) whether a verb really pertains to the relation in text or the result arose just by some strange incidence. For example, looking at our result table, s/he might ask if it is really typical (and thus worth modelling) for cities *to be known* for their museums. Display of the underlying text fragments (which are not overwhelmingly numerous in our case) would be of much help.

Acknowledgements

The work was partly carried out during M. Kavalec's stay at *Forschungszentrum Informatik* (FZI) Karlsruhe, Germany, supported by Marie Curie Fellowship, Semantic Web Center of Excellence, IST, MCFH-2001-00435. A. Maedche participated in the project during his appointment as Head of the Knowledge Management Group at FZI Karlsruhe. M. Kavalec and V. Svátek are partially supported by grant no.201/03/1318 of the Grant Agency of the Czech Republic.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. ACM SIGMOD Conference on Management of Data, 207–216, 1993.
2. 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.
3. Byrd, R., Ravin, Y.: Identifying and Extracting Relations in Text. In: Proceedings of NLDB 99, Klagenfurt, Austria, 1999.
4. Dill, S. et al.: SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In: Proc. WWW2003, Budapest 2003.
5. Faure, D., Nédellec, C.: ASIUM: Learning subcategorization frames and restrictions of selection. In: ECML'98, Workshop on Text Mining, 1998.
6. 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.
7. Kavalec, M., Svátek, V.: Information Extraction and Ontology Learning Guided by Web Directory. In: ECAI Workshop on NLP and ML for ontology engineering. Lyon 2002.
8. 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.
9. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer, 2002.
10. Maedche, A., Staab, S.: Mining Ontologies from Text. In: EKAW'2000, Juan-les-Pins, Springer, 2000.
11. 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.
12. Missikoff, M., Navigli, R., Velardi, P.: Integrated approach for Web ontology learning and engineering. IEEE Computer, November 2002.