Relation Labelling in Ontology Learning: Experiments with Semantically Tagged Corpus

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
{kavalec,svatek}@vse.cz

Abstract. Ontology learning from text can be viewed as auxilliary technology for knowledge management application design. We proposed a technique for extraction of lexical entries that may give cue in assigning semantic labels to otherwise 'anonymous' non–taxonomic relations. In this paper we present experiments on semantically annotated corpus *SemCor*, and compare them with previous experiments on plain texts.

1 Introduction

Ontologies are the backbone of the prospective semantic web as well as of a growing number of knowledge management systems. Recently, *ontology learning* (OL) has been proposed to overcome the bottleneck of their manual development. It relies on combination of shallow text analysis, data mining and knowledge modelling. In [9], three subtasks of OL have been systematically examined: lexical entry extraction (also viewed as concept extraction), taxonomy extraction, and *non–taxonomic relation extraction* (NTRE), considered as most difficult. For example, the NTRE component [10] of the *Text–to–Onto* tool [11] 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') even if they know that *some* relation between them has evidence in data. 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. Finally, 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 entries* picked up from relevant texts thus may give an important cue.

In our previous work [8] we experimented with documents from the Lonely Planet website[1] describing world locations. The *Text-to-Onto* tool was used for concept extraction and NTRE (components previously build by Maedche et al.

[1] http://www.lonelyplanet.com/destinations

[11, 10]) as well as for suggestion of *relation labels* (the newly added component). There is agreement in the NLP community that relational information is often conveyed by *verbs*; our technique thus selects verbs[2] (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.

This experiment[3] was relatively successful in discovering verbs that were (by human judgement) relevant labels for the given concept pairs. The *precision* was acceptable, since relevant verbs were mostly separated from irrelevant ones. The 12 triples suggested as most appropriate by the system typically corresponded to topo–mereological relations. For example: an island or country may be *located* a world–geographical region, a country may *be a country* of a particular continent and may be *located* on an island or *consist of* several islands[4], a city may *be home of* a famous museum etc. A weak point however was the low *recall* (wrt. the 5MB corpus), which we partly attribute to the following:

- The authors expressed the same information in many different ways, which makes the Lonely Planet data lexically very *sparse*.
- Ambiguous words were assigned all possible meanings; this of course added *noise* to the data and still decreased the chance for coherent results.
- Relation extraction was superposed over (automated) *concept extraction*; results of the former were negatively influenced by the flaws of the latter.

To eliminate some drawbacks of the first experiment, we adopted a semantically annotated text corpus named *SemCor*. Section 2 describes our approach to suggesting relation labels, section 3 presents the new experiments with *SemCor*, section 4 reviews related work, and section 5 wraps up the paper.

2 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*.

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 entries corresponding to associated concepts[5].

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.*

[2] Identified using a part-of-speech (POS) tagger.
[3] For more detail on the Lonely Planet experiment see [8].
[4] Example of *multiple relations* between the same concepts.
[5] We currently ignore the *ordering* of verb and concepts, to reduce data sparseness.

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 is conditional frequency (empirical probability)

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

where $|.|$ denotes set cardinality, and t_i are the VCC(n)–transactions. 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 entries corresponding to the concept 'city' often occurred together with the verb 'to reach', and the same held for lexical entries 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 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:

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

(the meaning of $P(c_1/v)$ and $P(c_2/v)$ being obvious). In the Lonely Planet experiment discussed above, the threshold value of $AE(c_1 \wedge c_2/v)$ for discrimination of relevant verbs from irrelevant ones was about 1.5.

3 Experiments

In order to overcome some difficulties mentioned above, we adopted *SemCor*[6]: a part of Brown corpus[7] semantically tagged with WordNet[8] senses. All open word classes (nouns, verbs, adjectives and adverbs) are tagged in 186 documents, with 2.18 MB overall. Advantages over an ad hoc document collection such as Lonely Planet immediately follow from reduced ambiguity:

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 do without a POS tagger, which also exhibited significant error rate.

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

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*[9]. We identified each of them with the WordNet synset containing the word sense person#1 (or group#1 or location#1, respectively). Any word tagged with 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[10]. The corpus contains 47701 sense-tagged verb occurrences[11]. In all three experiments below, we set the maximal verb-to-concept distance (n) to 5.

In the first experiment with *SemCor* we grouped the verbs directly by the *synset* they belong to; this yielded 4894 synsets. Table 1 shows the top synsets according to the *AE* score, for the Person-Group concept pair[12]. In the second experiment we generalised each verb by taking its (first-level) *hypernym synset*; we obtained 1767 synsets. Top ones for the same concept pair are in Table 2. In the third experiment we attempted to introduce some 'domain bias' through separately processed 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; however, only a fraction of verbs suggested as labels for a particular relation was indeed relevant. This was obviously due to data sparseness, even in the hypernym synset setting.

In the first two experiments, the quality of results was comparable to the Lonely Planet experiment despite the smaller and broader corpus. Most verbs with high AE measure seem to be potential labels for relations between Person and Group (and similarly for the other two concept pairs not shown here). This supports the hypothesis that our method could provide useful hints for an ontology designer. Human effort is of course still needed to filter out incidental results or e.g. to handle semantically incomplete expressions such as 'act as'.

In some cases, the impact of *verb generalisation* seems positive. For example, 'hire' (as definitely an important label) only was on 18th position in the verb synset version, while it floated up in the verb hypernym version. On the other hand, generalising may sometimes obscure the original meaning, e.g. the 'serve, function' synset is result of generalisation of 'act as' (the latter being probably more characteristic for Person-to-Group relationship). Sometimes even the one level of WordNet hypernymy may lead to overly general meaning, e.g. 'form, organize' is generalised to 'make, create', which scores much lower and thus does not appear among the top candidates. It seems that a combination of verbs directly found in text and of their careful generalisations might be the best blend to be presented to ontology designer.

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

[10] In Lonely Planet experiment we had 157 concepts with about 70000 occurrences.

[11] About 75000 verb occurrences were identified by the POS tagger in Lonely Planet.

[12] 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 Person and Group.

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 1. Suggested relations between Person and Group – verb synset version

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 design of proper relations is a creative task: 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 [7] 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 [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 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 [3] 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 [6] synergistically builds two hierarchies: that of concepts and that of verb sub-categorisation 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.

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
unite, unify	7	2.38
direct	21	2.14
announce, denote	9	2.06
appoint, charge	5	2.03
denounce	7	2.03
arrive, get, come	23	2.01
note, observe, mention, remark	9	1.94
hire, engage, employ	12	1.93
promote, upgrade, advance, kick upstairs, raise, elevate	5	1.93
re-create	11	1.81
join, fall in, get together	15	1.80

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

Another stream, 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* [10], the labelling problem was left upon the ontology designer. The same holds about the NTRE component of *DODDLE* [13], which only differs by a more sophisticated way of transaction construction. In the *OntoLearn* project [12], WordNet mapping was used to automatically assign relations from a small predefined set (such as 'similar' or 'instrument'), not focusing on verbs.

Interesting is the *OntoLT* plug-in to Protégé [4], which does not distinguish OL 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 linguistically annotated by means of another automatic tool: 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. NTRE corresponds to slot creation; the lexical label for new slot is directly transferred from (a single occurrence of) linguistic predicate within the phrase on which a slot-creation rule is applied.

5 Conclusions and Future Work

The experiments suggest that referring to the *right sense of words* improves the quality of relation labelling, and so might do the *grouping of verbs* by their meaning. Although we usually lack precise senses of words in real-world settings, in knowledge management applications it is often possible to restrict the senses of words with respect to a narrow domain. In particular, polysemous verbs typically become monosemous in the context of domain-specific terms.

A problematic point of the method is obviously the *direct mapping* from co-ocurrences of terms onto 'deep' ontological relations. It improperly suggests e.g. verbs that typically occur in some larger semantic context involving (among other) the two concepts in question but does not correspond to immediate relation between them at all. In the future, we thus plan to make the method more *linguistic-aware*, for example, to employ a chunker 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 eventually migrate to a *domain-specific* collection of texts relevant to a knowledge management application; this is crucial for determining the real value of our approach.

The research is partially supported by grant no.201/03/1318 of the Czech Science Foundation. Initial part of the work was carried out during M. Kavalec's stay at FZI Karlsruhe, Germany, in collaboration with Alex Maedche.

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. Faure, D., Nédellec, C.: ASIUM: Learning subcategorization frames and restrictions of selection. In: ECML'98, Workshop on Text Mining, 1998.
7. 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.
8. Kavalec, M., Maedche, A., Svátek, V.: Discovery of Lexical Entries for Non-Taxonomic Relations in Ontology Learning, In: Van Emde Boas, P., Pokorný, J., Bieliková, M., Štuller, J. (eds.). SOFSEM 2004. Springer LNCS, 2004, 249–256.
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.
13. Sugiura, N., Shigeta, Y., Fukuta, N., Izumi, N., Yamaguchi, T.: Towards On-the-Fly Ontology Construction – Focusing on Ontology Quality Improvement. In: Proc. ESWS-04, Heraklion 2004.