Information Extraction and Ontology Learning Guided by Web Directory

Martin Kavalec[1] and **Vojtěch Svátek**[2]

Abstract. The paper presents our ongoing effort to create an information extraction tool for collecting general information on products and services from the free text of commercial web pages. A promising approach is that of combining information extraction with ontologies. Ontologies can improve the quality of information extraction and, on the other hand, the extracted information can be used to improve and extend the ontology. We describe the way we use Open Directory as training data, analyse this resource from the ontological point of view, present some preliminary results related to information extraction, and outline our plans for building and deploying the ontology.

1 INTRODUCTION

Lack of explicit semantics and, consequently, poor machine understandability are commonly known problems of the current World Wide Web. In order to excavate *implicit* semantics from the full text of web pages, we can take advantage of both:

- Collections of operational *extraction patterns* (most often, in the form of rules) that specify at which points in the stream of (marked-up) text valuable information should be taken over. The nature of the patterns can be linguistic or surface-form-based (e.g. regular expressions).
- *Ontologies of problem domains* consisting of both the conceptual and lexical part. The identification of lexical items in the text leads to the abstraction of generic concepts, which can, in turn, be used as classes for extracted textual metadata characterising the web pages.

The dividing line between the extraction patterns and lexical ontologies is not always clear; we can roughly distinguish the patterns as being (to some extent) structural and having a lower degree of domain dependency.

A promising approach is that of combining information extraction with ontologies. Ontologies can improve the quality of information extraction and, on the other hand, the extracted information can be used to improve and extend the ontology, see [6]. A common strategy for this process is *bootstrapping*: a certain amount of manually labelled training data is initially provided, which serves for iterative labelling of unseen data associated via some properties with the original data. We however assume that the amount of manual labelling can be further restricted via the *reuse of public resources* with similar content and structure as the target knowledge.

[1] Department of Information and Knowledge Engineering, University of Economics, Prague, 13067 Praha 3, Czech Republic, e-mail: kavalec@vse.cz
[2] Department of Information and Knowledge Engineering, University of Economics, Prague, 13067 Praha 3, Czech Republic, e-mail: svatek@vse.cz

The goal of our effort described in this paper is relatively modest: to extract information about (mostly generic) *products, services and areas of competence of companies*, from the free text chunks embedded in web presentations.[3] For this sort of information, an abundant reusable resource are web directories such as Yahoo! or Open Directory. We have based our experiments on the 'Business' branch of *Open Directory* (`http://dmoz.org`). Both the hierarchy of the *directory headings* and the categorization of *links* listed in each node are valuable sources of information. From the categorization of web links we can obtain *labelled* training data for information extraction, while the hierarchy could be used as source for building a (lightweighted) ontology of the domain corresponding to the given branch.

Section 2 describes the method of using the directory to acquire labelled examples for information extraction and shows some preliminary results of subsequent learning. Section 3 outlines the way how the results of learning can be exploited in a distributed architecture for web analysis. In section 4 we analyse the structure of the Open Directory headings, in section 6 our approach is compared to some other projects and in section 7 we summarize our plans for the future.

2 MINING INDICATOR TERMS THROUGH DIRECTORY HEADINGS

The general description of the company profile, area of competence, products and services is usually not too extensive but stylistically well-formed. This favourises the use of deeper *linguistic* techniques, in contrast to surface techniques (such as regular-expression-based), which are often used for information extraction from idiosyncratic, abridged documents (e.g. advertisements or medical records).

Our assumption is that the *directory headings* (such as `.../Manufacturing/Materials/Metals/Steel/...`) coincide with the generic names of products and services—let us nickname them *informative terms* in this paper—offered by the owners of the pages referenced by the respective directory page. By matching the headings with the page full texts, we obtain sentences that contain the informative terms. The terms situated near the informative terms in the syntactical structure of the sentence are candidates for *indicator terms*, provided they occur frequently on pages from various domains. The resulting collection of indicator terms

[3] Currently, we do not consider other company information such as cooperation with other companies or financial results, which is much sparsely present in common web pages. We also ignore the possibility to extract company information (as a specific sort of web page metadata) from the micro-level structures of *HTML mark-up*, which is the subject of a project running in parallel.

can be, conversely, the basis of extraction patterns for discovering informative terms in previously unseen pages.

The knowledge asset embedded in web directories is the judgement of human indexers who have assigned the pages under the particular heading(s). Naturally, informative terms on the page need not always correspond to the existing directory headings, e.g. due to synonymy. As consequence, our method will extract (without the help of a thesaurus) only a fraction of the sentences with informative terms. This however does not disqualify the method, since, in this training phase, we aim at discovering indicator terms rather than at identifying the informative terms themselves. The small degree of completeness of the method is actually compensated by the hugeness of the material available[4] in the directories. Namely, the 'Business' subhierarchy of Open Directory that we have exploited in our experiments points to approx. 150,000 pages overall, each of these containing the 'heading' terms (from the referencing node or one of its ancestors) in two sentences, on the average.

We have tested the training phase of our method on a sample of 14,500 sentences[5] containing the 'heading' terms. The syntactical analysis has been carried out using the free *Link Grammar Parser*[6] [10]. Our working hypothesis was that the abovementioned indicative function is, in most cases, conveyed by *verbs* (and verb phrases). Therefore, in the initial experiments, the verbs that occurred the closest (in the parse tree) to informative terms have been counted, arranged into a frequency table, and ordered by ratio of their relative frequency of occurrence near some informative term to their relative frequency in general. Eight[7] most promising verbs have been chosen for the experimental collection. Most of these are likely to be associated with informative terms, e.g. 'our assortment *includes*…', 'we *manufacture*…', 'in our shop you can *buy*…'.

The method itself has been described in more detail in [4]; by now, preliminary testing results are also available, see Tab. 1. For the test, 130 sentences containing some indicators were randomly selected and each of them was *manually labelled*. The labelling amounted to the subjective estimation whether the sentence contains the target informative terms or not. This is sometimes difficult—e.g. due to missing context, special terminology and domain specific product names; see for example the sentence:

> We are equipped to run any grade of corrugated from E-flute to Triplewall, including all government grades.

Therefore, some unclear sentences were labelled with '?' and then counted once as negative and once as positive test cases. Some sentences contained the company name but no usable information on the products, e.g.

> Industrial Metals Inc. is committed to provide you with exceptional service.

Although named entities are often valued in the information extraction field, we considered these sentences as negative test cases, too, since we focus on *generic* names of products/services or of their providers. The testing results (including ad-hoc inspections not covered by the presented table) suggest that some general[8] verbs–such as 'use' or 'include'–need to be extended to more complex *phrases*, possibly again via selecting the neighbouring terms with frequent occurrence. Also, clearly, certain nouns and noun phrases could play the role of indicators, too.

Table 1. Test of the indicative verbs

indicator	−	?	+	precision
include	8	4	18	60–73%
provide	9	3	28	70–78%
offer	6	1	21	75–79%
specialize	0	1	18	95–100%
(other)	3	5	5	38–77%
total	26	14	90	77–80%

Due to the tedium of the abovementioned manual labelling, we are not able to measure directly the *coverage* of a collection of indicators: this would amount to considering the full set of sentences in the selected sample of web pages. An indirect measure of coverage, which can be obtained automatically, is the number of pages in the sample that contain one or more indicators from the collection. On the pages directly referenced by directory nodes, this measure was rather low, between 10-20%; however, if we manually pre-filtered out pages with no or minimal free-text content (such as intro or menu pages), the proportion increased to 70-80%: the fact that this result was obtained for a collection of *eight* indicators suggests that the cross-domain variability of these terms might be relatively limited. Note that, even if a set of indicators could not directly be used, due to low coverage, for systematic filling of information extraction templates, it could still be acceptable for the discovery of new terms for the *ontology of products and services*, see section 5.

3 INTEGRATION OF INDICATOR-BASED ANALYSIS INTO A MODULAR ARCHITECTURE

Indicator-based linguistic analysis, as described in this paper, has only limited capabilities with respect to the heterogeneous content of commercial web pages. In order to bring useful results, it is thus being integrated into a modular architecture currently under development. The central idea of the architecture, named *Rainbow*[9] [12] (Reusable Architecture for INtelligent Brokering Of Web information access) is the separation of different web analysis tasks according to the *syntactical type of data* involved. Communication within *Rainbow* is based on the simple *SOAP* [1] communication protocol. Services provided by the individual modules – acquisition of data from the web, conversion to well-formed XML, different forms of semantic analysis of data and, finally, visualisation of results – are described by means of *WSDL*, the Web Service Description Language [3]. Indicator-based linguistic analysis, as described in this section, has been implemented as one of the web services within the first prototype of *Rainbow*, currently in the form of sentence extraction. The 'interesting sentences' are part of the output of the visualisation component, which can be installed as a plug-in panel of the

[4] As we dispense with manual labelling, processing a larger sample of data is merely the matter of computer time/storage.

[5] I.e. about 5% of the total of such sentences.

[6] The choice was motivated partly by the immediate availability of the parser, partly by the hypothesis that a linked-based parser could support the presumed 'navigation' over the dependency structures better than parsers based on constituent grammars.

[7] We hope to build a more comprehensive collection using a larger sample of pages, and possibly more domain-specific collections for sub-branches of 'Business'.

[8] Even the verb 'to be', which has no significance of its own, could presumably be the starting point for finding useful indicator phrases.

[9] Beyond the acronym, the name is motivated by the idea that the individual modules for analysis of web data should synergistically 'shed light' on the web content, in a similar way as the different colours of the rainbow join together to form the visible light.

Mozilla browser. In addition to linguistic analysis, *explicit metadata* (in META tags) are currently processed; moreover, *similar pages* are displayed thanks to the respective web service provided by Google.

For the next version of the architecture, an earlier-developed URL analyser [11] is being adapted; separate modules for the analysis of HTML structures, inline images, and link topology structures are also under design. Shared domain *ontologies* will serve for verification of semantic consistency of web services provided within the distributed system. Clearly, an advanced version of the architecture should be able to overcome the mentioned problem of directory links pointing to the 'barren' pages of the particular website: analysis of keywords and HTML structures on the start-up pages, as well as of the URLs of embedded links, will navigate the proper metadata extractor towards the most promising pages or page sections. Such parts of company websites, named e.g. *about-us*, *profile* etc., are quite common and usually contain larger segments of syntactically correct text.

4 ONTOLOGICAL ANALYSIS OF WEB DIRECTORIES

Web directory hierarchies are sometimes mistaken for ontologies; however, as already observed by Uschold [13], they are rarely valid taxonomies. It is easy to see that subheadings are often not specializations of headings; some of them are even not *concepts* (names of entities) but *properties* that implicitly restrict the extension of a preceding concept in the hierarchy. Consider for example `.../Industries/Construction_and_Maintenance/Materials_and_Supplies/Masonry_and_Stone/Natural_Stone/International_Sources/Mexico`.

Semantic interpretation of a representative sample of directory paths has revealed that

- terms and phrases in individual headings belong to quite a small set of *classes*, and
- surface 'parent-child' arrangement of headings belonging to particular classes corresponds (with a certain degree of ambiguity) to 'deep' ontological *relations*.

The result of this effort was a *meta-ontology of directory headings* plus a collection of *interpretation rules*. The diagram at Fig. 1 depicts the essence[10] of the *meta-ontology*. Boxes correspond to classes, full edges to named relations, and dashed edges to the class-subclass relationship. Reflexive binary relations are listed inside the respective boxes. Examples of informally expressed *interpretation rules* are in Tab. 2.

5 INFORMATION EXTRACTION AND ONTOLOGY LEARNING

Plain indicator terms, gathered by means of the fully automated technique described in section 2, are by themselves powerful enough to extract *sentences* that are likely to contain *some kind of* interesting information about the company. We can even, in many cases, access this information thanks to simple heuristics over the parse-tree, such as:

[10] For better readability, we have e.g. omitted the notion of 'Location', which may also be important to extract but is not directly related to the commercial profile of the company.

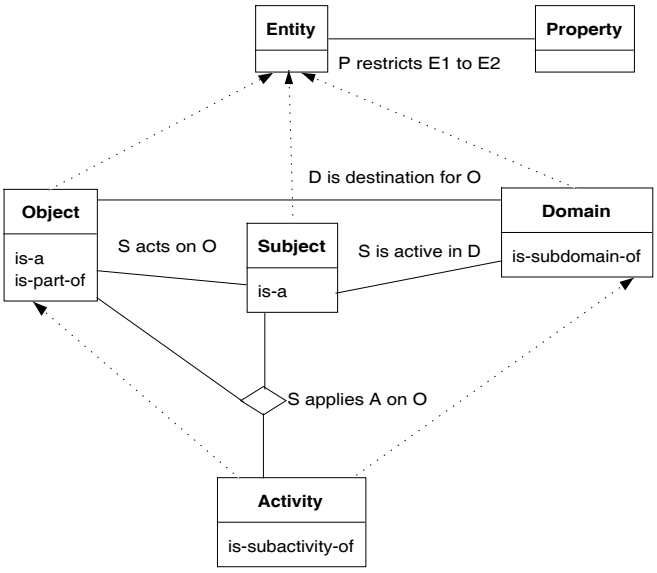

Figure 1. The ontology of web directory headings

Table 2. Examples of interpretation rules

Rule no.	Path pattern	Ontology relation
1	Subj/Prop	'Prop_Subj' *is-a* Subj (or, Prop *restricts* Subj *to* 'Prop_Subj')
2	Dom1/Dom2	Dom2 *is-part-of* Dom1
3	Obj1/Obj2	Obj2 *is-a* Obj1
4	Dom/Prop	'Prop_Dom' *is-part-of* Dom

Rule no.	Example
1	`Publishers/Academic_and_Technical`
2	`Security/National_Security`
3	`Electric_Motors/AC_Motors`
4	`Manufacturing/Electrical`

If the immediate object of the *indicator verb* is a generic *set-semantic expression* such as 'range of', 'family of', 'assortment of' etc. then output the *indirect attribute* of the object; otherwise output the *object* itself.

Universal extraction patterns however impose strong assumptions on the whole collection of indicators. A more sensitive method should take account of the *classes* of indicators/headings revealed by ontological analysis. If we learn the indicators for each class of information (such as 'subjects', 'objects' or 'domains') separately, we could be able to perform true *information extraction* in the sense of filling database templates. Conversely, if the informative terms thus discovered coincide with the headings of directory nodes referencing the particular page, we can automatically 'restore the identity' of these headings. With the help of generic interpretation rules such as those shown in Tab. 2, fragments of true taxonomies (possibly several interconnected ones, for 'subjects', 'objects'..., as specified by the meta-ontology) could be built. We can understand this as a two-step *ontology learning* process using two resources: text and the hierarchies of headings. Obviously, the result of this process will still be rather incomplete, and should be enhanced using other ontology-learning techniques, taking into account co-occurrences (and linguistic dependencies) of terms in the text beyond the headings.

These two tasks represent a *closed loop*: as soon as we have classified the headings, we can learn class-specific indicators[11]. From the other side: as soon as we have class specific indicators, we can use them for the classification of headings. Since the first step in this loop has to be done by a human, a more viable approach seems to be that one starting by *classifying the directory headings*. For this task we could use the WordNet lexical database. One reason for this are some regularities and similarities in the structure of Open Directory: some of the headings could thus be even classified semi-automatically with the help of heuristic rules. Another interesting possibility is to classify the headings by matching them to a generic lexical ontology such as *WordNet*.

6 RELATED WORK

The combination of information extraction and ontology learning has previously been described by Maedche [6]. The main novelty of our approach is the use of a public *web directory*.

Li, Zhang and Yu also use the Link parser and describe in [5] how to learn mapping from the link grammar to RDF statements. Their work shows advantages of link grammar over constituent grammar for this task and demonstrates feasibility of this task.

While directories have already been used for learning to classify *whole documents*, by Mladenic, [8], their use for *information extraction* seems to be innovative.

There is also some similarity to Brin [2], which targets on automated discovery of extraction patterns using *search engines*. The patterns can be used to find relations, such as books, i.e. pairs (author, title). However, the patterns are simply based on characters surrounding the occurrence of the investigated relation. In comparison, we aim at finding less structured information, for which such simple patterns wouldn't be sufficient.

Finally, the use of bootstrapping and other statistical methods for information extraction has also been presented e.g. in [7] and [9].

[11] The class specific indicators will apparently be more complex than the current ones.

7 FUTURE WORK

Given the three topics of the paper, actual results (based on Open Directory data) have been so far obtained only for indicator learning[12] and ontological analysis. The most challenging task that remains is the completion of the *information extraction & ontology learning loop*.

Since both these tasks can easily be related to the objectives of the *Semantic web*, we would also like, in longer run, to adapt the technique to the standards of usual *explicit* metadata. The information extracted can be, for example, forged to RDF triples, with indicator collections being accessible over the web.

ACKNOWLEDGEMENTS

The authors would like to express thanks to Jirka Kosek, main developer of the integrated *Rainbow* architecture that currently serves for evaluation of the partial techniques described in this paper.

The research has been partially supported by the grant no. 201/00/D045 (Knowledge model construction in connection with text documents) of the Grant Agency of the Czech Republic.

REFERENCES

[1] D. Box et al,*Simple Object Access Protocol (SOAP) 1.1*, W3C Note, 2000. `http://www.w3.org/TR/2000/NOTE-SOAP-20000508`

[2] S. Brin, *Extracting Patterns and Relations from the World Wide Web.* In WebDB Workshop at EDBT'98.

[3] E. Christensen et al, *Web Services Description Language (WSDL) 1.1*, W3C Note, 2001. `http://www.w3.org/TR/2001/NOTE-wsdl-20010315`

[4] M. Kavalec, V. Svátek and P. Strossa, *Web Directories as Training Data for Automated Metadata Extraction*. In: Semantic Web Mining, Workshop at ECML/PKDD2001, Freiburg 2001, 39-44.

[5] Y. Li, L. Zhang, Y. Yu, *Learning to Generate Semantic Annotation for Domain Specific Sentences*. In: K-CAP 2001 Workshop on Knowledge Markup & Semantic Annotation, October 21, 2001, Victoria B.C., Canada.

[6] A. Maedche, G. Neumann and S. Staab, *Bootstrapping an Ontology-Based Information Extraction System.* Studies in Fuzziness and Soft Computing, editor J. Kacprzyk. INTELLIGENT EXPLORATION OF THE WEB, P.S. Szczepaniak, J. Segovia, J. Kacprzyk, L.A. Zadeh, Springer 2002/01/01

[7] A. McCallum, K. Nigam, *Text Classification by Bootstrapping with Keywords, EM and Shrinkage.* In ACL'99 Workshop for Unsupervised Learning in NLP, 1999.

[8] D. Mladenic, *Turning Yahoo into an Automatic Web-Page Classifier*. In: Proc. 13th European Conference on Aritficial Intelligence, ECAI'98, 473-474.

[9] E. Riloff and R. Jones, *Learning Dictionaries of Information Extraction by Multi-Level Bootstrapping*. In Proc. 16th Nat. Conf. Artificial Intelligence (AAAI-99).

[10] D. Sleator and D. Temperley, *Parsing English with a Link Grammar.* In Third International Workshop on Parsing Technologies, August 1993.

[11] V. Svátek and P. Berka, *URL as starting point for WWW document categorisation*. In: RIAO'2000 – Content-Based Multimedia Information Access, Paris, 2000.

[12] V. Svátek, J. Kosek J. Bráza, M. Kavalec, J. Klemperer and P. Berka, *Framework and Tools for Multiway Extraction of Web Metadata.* Accepted for: Information Systems Modelling, Rožnov 2002.

[13] M. Uschold and R. Jasper, *A Framework for Understanding and Classifying Ontology Applications*. In: Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends.

[12] Even indicator learning could be improved in many ways, some of which are mentioned in the end of section 2.