Integrating Structural Data into Methods for Labeling Relations in Domain Ontologies

Gerhard Wohlgenannt
Institute for Information Business
Vienna Univ. of Economics
Vienna, Austria
Email: wohlg@ai.wu.ac.at

Albert Weichselbraun
Institute for Information Business
Vienna Univ. of Economics
Vienna, Austria
Email: aweichse@ai.wu.ac.at

Arno Scharl
Department of New Media Technology
MODUL Univ. Vienna
Vienna, Austria
Email: scharl@modul.ac.at

***Abstract*—This paper presents a method for integrating DBpedia data into an ontology learning system that automatically suggests labels for relations in domain ontologies based on large corpora of unstructured text. The method extracts and aggregates verb vectors for semantic relations identified in the corpus. It composes a knowledge base which consists of (i) centroids for known relations between domain concepts, (ii) mappings between concept pairs and the types of known relations, and (iii) ontological knowledge retrieved from DBpedia. Refining similarities between the verb centroids of labeled and unlabeled relations by means of including domain and range constraints applying DBpedia data yields relation type suggestions. A formal evaluation compares the accuracy and average ranking performance of this hybrid method with the performance of methods that solely rely on corpus data and those that are only based on reasoning and external data sources.**

***Keywords*-ontology learning; structural data; data integration; relation labeling**

I. Introduction

Ontologies formally specify a conceptualization of an application domain (1) and therefore provide the means for a common understanding of domain concepts and relations among different stakeholder groups. When domains evolve, there is a constant need to update and refine domain-specific ontologies to ensure their usefulness. The bottleneck and cost-driver in ontology learning tends to be the availability of expertise and qualified human resources. Automated approaches address this problem by supporting ontology engineers, improving their productivity, and reducing the human input required.

Identifying and labeling non-taxonomic relations are among the ontology learning subtasks that are considered most challenging (2). Events announcing competitions such as the Fourth International Workshop on Semantic Evaluations (SemEval 2007, nlp.cs.swarthmore.edu/semeval) underscore the growing importance of identifying semantic relations. This paper distinguishes between methods applying (i) *corpus analysis*, extracting information from corpus resources; (ii) *corpus enrichment*, extending and annotating the corpus via external resources such as Wikipedia or Google; and (iii) *semantic inference and validation*, incorporating data from Semantic Web sources and investigating relations by reasoning upon this data.

Corpus analysis applies linguistic patterns (3; 4; 5), association rules (6), kernel-based approaches (7) and other techniques from the fields of artificial intelligence, statistics, mathematics, and combined approaches to the problem of relation type discovery.

Corpus enrichment integrates external resources to increase the accuracy of the relationship labeling. Sanchez and Moreno (8) present an approach using verbs from sentences containing domain concepts and search engine queries for relationship labeling. Giuliano et al. (9) use WordNet synsets and hypernym relations to refine kernel methods for extracting semantic relations.

Semantic Inference and Validation integrates structural data from semantic Web resources, a method that has become quite popular in recent years. Scarlet and Watson (10) leverage ontological knowledge from one or multiple ontologies to determine the relation between pairs of concepts. Lehmann et al. (11) query structural data from DBpedia (www.DBpedia.org) to identity relations between concepts by finding paths between these concepts.

Despite the potential of the approaches presented above, their usefulness is limited by the so-called *knowledge acquisition bottleneck* (12), a term that refers to the difficulty of creating and maintaining extensive knowledge bases. To overcome the restrictions imposed by the knowledge acquisition bottleneck, the approach presented in this paper combines reasoning based on external structural data with machine learning methods.

The remainder of this paper is structured as follows: Section II presents the relation type labeling component and elaborates on the integration of DBpedia into the identification process. Section III evaluates the component using different experimental setups. The paper concludes with a summary and outlook in Section IV.

1529-4188/09 $25.00 © 2009 IEEE
DOI 10.1109/DEXA.2009.26

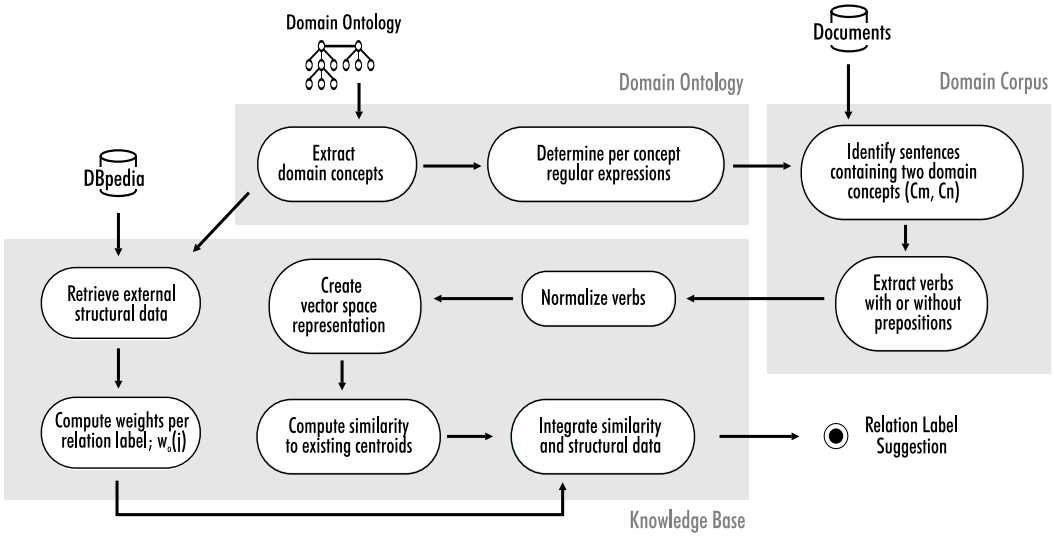

Figure 1. Architecture of the relation type labeling component.

II. METHOD

The method presented in this section suggests labels for unlabeled relations in domain ontologies. It is independent from any particular ontology learning system, but has been developed as a compontent of the framework introduced by Liu et al. (4), which extracts domain terminology, taxonomic, and unlabeled relations from text corpora.

Figure 1 illustrates the labeling process. Based on known relations and an input corpus, the framework extracts verbs from sentences containing the domain concepts (C_m, C_n) participating in the relation l_{mn} and stores this data in its knowledge base. The similarity between verb vectors of an unlabeled relation l^*_{mn} and the data in the knowledge base yields a similarity score between known labels and the unlabeled relations. Querying DBpedia via SPARQL maps domain concepts to types such as `Person`, `Organization`, and `Topic`. Matching this information with domain and range constraints for the suggested label allows removing invalid relation type labels or decreasing their similarity score. Finally, the component selects the label with the highest score for the unlabeled relation.

A. Composing Verb Vectors

The method applied in this research is based on corpus analysis algorithms developed by Weichselbraun et al. (13)

The relation label suggestion component uses machine learning techniques and ontological knowledge retrieved from external resources to compile a knowledge base (KB). Consulting this knowledge base yields suggestions for the relation types of unlabeled relations. In the following, we provide a formal description of this matching process.

Each term (C) in the domain ontology is represented by a list of regular expressions (C^r) and connected to other concepts by links $l_{mn}(C_m, C_n)$. Equation 1 defines the list of verb vectors L^v_{mn} that characterize the semantic relation between the concepts C_m and C_n.

$$L^v_{mn} = \{verbs(s_i) \quad | \quad \textit{match}(C^r_m, s_i) \wedge \textit{match}(C^r_n, s_i) \quad (1)$$
$$\wedge \quad \textit{idx}(C^r_m, s_i) < \textit{idx}(C^r_n, s_i) \quad \}$$

L^v_{mn} is composed of the vector space representation $\vec{v}_i := verbs(s_i)$ of verbs occurring in a sentence s_i together with the domain concepts C_m and C_n. The *match* operator returns true if sentence s_i matches at least one of the regular expressions in the list C^r.

The *verbs(s_i)* operator returns a vector space representation of the infinitive form of all verbs present in sentence s_i. In some cases, the use of prepositions changes the direction or even the meaning of relations (e.g. `deal in` versus `deal with`). For assessing the effect of prepositions on the method's accuracy, we compiled two knowledge bases (KB, KB') that support two different *verbs(s_i)* functions. One knowledge base (KB) solely considers verbs, the other one (KB') stores verbs and prepositions (if available) for the suggestion process. The evaluation in Section III provides a comparison of the average ranking performance of relation type labels computed with these approaches.

The order of the concepts is important for the evaluation process. We define that $l_{mn}(C_m, C_n) := \neg l_{nm}(C_n, C_m)$, which effectively reverses the direction of a relation. The *idx* operator in the second term of the definition ensures that the first concept (C_m) occurs before the second concept (C_n).

Equation 2 computes the centroid $\vec{V}_{mn}$, which represents the verb vector for the relation l_{mn} between the two concepts C_m, C_n.

$$\vec{V}_{mn} \quad = \quad \sum_{i=1}^{|L^v_{mn}|} \vec{v}_i / \Big| \sum_{i=1}^{|L^v_{mn}|} \vec{v}_i \Big| \qquad (2)$$

The knowledge base of the framework consists of (i) a list of all centroids $\vec{V}_{m_i n_j}$ representing the relation $l_{m_i n_j}$, (ii) the mapping $M_{mn \rightarrow j}$ assigning a label j to the relation $l_{m_i n_j}$, and (iii) the set of ontology snippets $\{O_1, O_2, ...O_n\}$ retrieved from external sources, containing some formalized knowledge about the domain.

$$KB = (\{\vec{V}_{m_1 n_1}, ...\vec{V}_{m_k n_l}\}, M_{mn \rightarrow j}, \{O_1, ...O_n\}) \quad (3)$$

B. Labeling Relations and Integrating Feedback

Applying Equation 1 and 2 yields the centroid $\vec{V}_{m^* n^*}$ for the unlabeled relation. The first step in determining the relation's type is computing the similarity between this centroid and all known centroids in the knowledge base using a similarity function (sim) by applying Equation 4:

$$s_{mn} = w_{o,m^* n^*}(\underbrace{M_{mn \rightarrow j}(mn)}_{j}) \cdot sim(\vec{V}_{m^* n^*}, \vec{V}_{mn}) \quad (4)$$

The current architecture uses the cosine measure as similarity function. The factor w_{o,m^*,n^*} considers domain knowledge using the following heuristic:

$$w_{o,m^* n^*}(j) = \begin{cases} 1.0 & \text{if } O_i \models C_{m^*} \in dom(j) \quad \wedge \\ & \quad O_k \models C_{n^*} \in range(j) \\ 0.01 & \text{if } O_i \models C_{m^*} \ni dom(j) \quad \vee \\ & \quad C_{n^*} \ni range(j) \\ 0.8 & \text{if } O_i \models C_{m^*} \in dom(j) \quad \vee \\ & \quad C_{n^*} \in range(j) \\ 0.5 & \text{otherwise.} \end{cases} \quad (5)$$

Equation 5 determines the weighting factor $w_{o,m^* n^*}$ based on whether the ontology implies ($\models$) the domain and/or range restrictions from the ontology or not. In cases where the component can not verify any domain and range restrictions (relations with no domain and range constraints; concepts for which no type could be identified), a weight of 0.5 is assigned. From the computed similarity s_{mn} and the mapping $M_{mn \rightarrow j}$ we determine a list of triples, containing the relation l_{mn}, the matching relation label j, and its similarity to the unlabeled relation $l_{m^* n^*}$.

The relation label j is determined using one of the following strategies: (i) selecting the relation label j of the relation l_{mn} with the highest similarity s_{mn}, (ii) computing the average of the similarity measures s_{mn} for each relation label j and selecting the label with the highest average similarity, or (iii) determining the average of the highest 30% of the similarity measures for each relation label j, and selecting the label corresponding to the highest average.

Domain experts either confirm or discard the suggested relation. This feedback is incorporated by adding the mapping $mn \rightarrow j$ to the mapping $M_{mn \rightarrow j}$ in the knowledge base. Such feedback therefore refines the knowledge base and constantly improves the component's accuracy.

III. EVALUATION

This section summarizes a series of experiments conducted to evaluate the performance of the outlined method based on 310 relations in the climate change domain. These 310 relations are built from 155 basic relations, adding relations with the concepts in reverse order. As this paper focuses on relation type detection, we decided to test and evaluate the method with a sufficient number of high-quality relations garnered by letting domain experts manually extend relation sets identified by the webLyzard ontology extension architecture (4).

The vector space representations in the evaluations consider verbs appearing in (i) the same sentence as the concepts (C_m, C_n), and (ii) within a sliding window size of five and seven words. Contrasting experiments considering prepositions with computations neglecting them allows assessing the influence of prepositions on the performance of the relation type suggestion component.

A. Experimental Setup

In the evaluation, we drew upon a list of 156 news media sites from the Newslink.org, Kidon.com and ABYZNewsLinks.com directories. The webLyzard suite of Web mining tools (www.weblyzard.com) crawled these sites and gathered about 200,000 documents per week. A domain detection service based on regular expressions helped compile an extensive domain-specific corpus with documents published between December 2008 and February 2009. A separate corpus assembled from environmental blogs complements the news media data.

Table I lists the relation types used for labeling relations and the number of sentences in the corpora satisfying Equation 1 (see Section II) from which verb vectors for that particular relation type could be extracted.

We used a total of 95,733 sentences from the corpus for evaluating the method, 56,634 of which were unique. The 310 relations from the test ontology were randomly split into training and testing sets of equal size.

linkType	¬ linkType	$\text{sentences}_{unique}$
subClassOf	superClassOf	4273
use	usedBy	12905
study	studiedBy	14807
hasEffectOn	isAffectedBy	22337
disjointWith	disjointWith	2312

Table I
RELATION TYPES USED IN THE EVALUATION

For all concepts appearing in the test relations, we tried to determine their *type* by querying DBpedia as illustrated in Table II. This information is used when applying *domain* and *range* restrictions. The system could determine the type of 58 out of 97 concepts. Concepts for which no type could be discovered are labeled as "unknown", and treated as proposed in Equation 5.

concept	type
Al Gore, scientist	person
NOAA, IPCC, OPEC	organization
fossil fuel, ecosystem	object topic
exploitation, peak oil	abstract topic

Table II
CONCEPTS AND THEIR RESPECTIVE TYPES

	verbs only		verbs and prepositions	
	sliding[1]	sentence	sliding[3]	sentence
nodir DBP	1.90 (1.87)	2.25 (2.19)	1.88 (1.83)	2.13 (2.08)
nodir VSM	2.33 (2.31)	2.86 (2.83)	2.28 (2.31)	2.69 (2.66)
dir DBP	2.58 (2.55)	2.73 (2.66)	2.50 (2.45)	2.68 (2.63)
dir VSM	3.21 (3.19)	3.57 (3.54)	3.25 (3.22)	3.64 (3.63)

Table III
AVERAGE RANKING PRECISION (ARP)

For each non-directed relation type, the knowledge base was trained with a set of 40-56 pre-defined concept-relation patterns. The number of verbs extracted from the corpora ranged from 21 to 7668 per training relation, depending on the extraction mode (whole sentence, sliding window). The average number was 684.

B. Results

In the experiments, the relation type suggestion component assigned an ordered list of distinct relation types presented in Table I to unlabeled relations. The evaluation distinguishes between suggestions derived from the vector space model (corpus analysis), and suggestions combining this model with semantic inference and validation based on DBpedia. The results for two configurations are presented as follows: Vector Space Model only (VSM) and Vector Space Model plus DBpedia (DBP). For rows marked with "dir", the relation type *and* direction were computed. Rows identified by the term "nodir" only consider the correct relation type for the evaluation. Table III summarizes the different approaches' ARP, specifying the average number of tries required to pick the correct relation type label from an ordered list of suggestions (the table contrasts computations based on a sliding window size of seven words with results computed with whole sentences). The average ranking precision (ARP) for randomly chosen relation types is 3.0 for guessing the correct label and 5.0 for picking the right label and direction. This measure is highly relevant, as the ontology relation type suggestion has been designed to aid domain experts in assigning relation types and indicates how many choices the domain expert has to check on average to identify the correct label. In order to precisely evaluate the performance gains that information from structural sources provide, we conducted a second evaluation restricted to the set of testing relations for which at least one concept "type" could be extracted. The results for those 148 relations are given in parentheses. As DBpedia data is extracted from Wikipedia automatically it is incomplete and not always correct, these problems are addressed in future research.

The ARP results show that the combined approach – with semantic validation (DBP) – clearly outperforms the VSM-only method. Applying Scarlet (scarlet.open.ac.uk), a method solely based on querying Semantic Web resources, to the evaluation task only yielded relation types for eight out of 155 testing relations. This is attributable to the knowledge acquisition bottleneck discussed in the introduction. Four out of eight relations were labeled correctly by Scarlet. We also encountered a case, in which Scarlet inaccuratly labeled relations due to an incorrect subClassOf relation in an external ontology (*oil subClassOf industry*) as described by d'Aquin et al. (12). Currently Scarlet does not influence the evaluation results significantly, so it is not included in Tables III and IV. Nevertheless, with the growth of the Semantic Web, we expect the number of found relations to rise dramatically, making it worthwhile to integrate Scarlet into the presented framework.

Using verbs extracted with sliding windows yielded better results than verbs from whole sentences. This is caused by the fact that sliding windows are more precise in respect to returning only verbs in the vicinity of concepts.

Table IV summarizes the results as a percentage of correctly identified relation types. The "1st guess correct" column shows the percentage of relations correctly identified by the first suggestion. The "2nd guess" column gives the percentage of relations correctly labeled by the first or second suggestion.

	1st guess correct (%)		2nd guess correct (%)	
	sliding	sentence	sliding	sentence
nodir DBP	65.1 (65.8)	64.5 (64.9)	79.7 (80.0)	74.8 (75.7)
nodir VSM	49.1 (50.0)	47.7 (47.3)	67.5 (67.5)	60.0 (60.1)
dir DBP	44.7 (45.8)	45.7 (48.6)	71.5 (71.7)	64.5 (66.2)
dir VSM	33.3 (34.1)	29.0 (29.1)	56.1 (55.8)	47.1 (47.9)

Table IV
CORRECTLY IDENTIFIED RELATION TYPES IN THE EVALUATION (SLIDING WINDOW SIZE OF SEVEN WORDS)

It is obviously much harder to guess the correct relation type and direction, where we have nine possibilities and a probability of about 11% when guessing randomly (compare Table I – the `disjointWith` relation type is symmetric), than guessing only the relation type, where there are five possibilities and a 25% chance of randomly guessing the correct label.

Conducting a Chi-squared test on the results presented in Table IV shows that the significance levels of the presented method exceed 99.99%. The accuracy of 65.1% for determining the correct label at the first guess (79.7% for second guess) in Table IV is equivalent to an F-measure of 0.79 (0.89) when retrieving relation types only.

One would expect that the proposed methods perform differently depending on the relation type, and experiments

confirm that intuition. The approach worked best for the relation type *study* with more than 80% of correct suggestions at the first guesses, and an ARP around 1.5. *Study* is particularly well suited, as it has a very clearly defined subject domain ('person', 'organization') and object range ('object topic', 'abstract topic'). The *disjointWith* relation type, by contrast, is hard to grasp with our approach with an ARP for relation types and direction of 3.5.

IV. Conclusions

This paper elaborates on the use of structural data from external resources to suggest labels for unlabeled relations based on classical corpus analysis methods. A method integrating a machine learning technique based on the vector space model with structural data from DBpedia is presented and evaluated. The main contributions of this research are: (i) introducing a novel approach that integrates structural data with a machine learning approach based on a vector space model for suggesting ontology relation type labels; (ii) presenting an extensive evaluation to assess the method's performance; (iii) outlining the advantages of combined approaches and (current) problems with methods solely relying on structural data.

Future research should emphasize the integration of additional, heterogeneous data sources - including strategies for resolving conflicts between annotations from multiple sources. Disambiguation and mediation techniques are a cornerstone for addressing this challenge and providing a more fine-grained and accurate assessment of concept types and therefore relation labels.

V. Acknowledgment

The project results have been developed in the IDIOM (Information Diffusion across Interactive Online Media; www.idiom.at) project funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG). The authors would like to thank Arinya Eller for proofreading the manuscript.

References

[1] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *International Journal of Human-Computer Studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[2] M. Kavalec and P. Spyns, *Ontology Learning from Text*. Amsterdam: IOS Press, 2005, ch. A Study on Automated Relation Labelling in Ontology Learning, pp. 44–58.

[3] P. Cimiano and J. Wenderoth, "Automatically learning qualia structures from the web," in *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 28–37.

[4] W. Liu, A. Weichselbraun, A. Scharl, and E. Chang, "Semi-automatic ontology extension using spreading activation," *Journal of Universal Knowledge Management*, vol. 0, no. 1, pp. 50–58, 2005, http://www.jukm.org/jukm_ 0_ 1/semi_ automatic_ ontology_ extension.

[5] M. Poesio and A. Almuhareb, "Identifying concept attributes using a classifier," in *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 18–27.

[6] A. Maedche, V. Pekar, and S. Staab, "Ontology learning part one - on discoverying taxonomic relations from the web," in *Web Intelligence*, N. Zhong, J. Liu, and Y. Yao, Eds. Springer, 2002, pp. 301–322.

[7] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *J. Mach. Learn. Res.*, vol. 3, pp. 1083–1106, 2003.

[8] D. Sánchez and A. Moreno, "Learning non-taxonomic relationships from web documents for domain ontology construction," *Data Knowl. Eng.*, vol. 64, no. 3, pp. 600–623, 2008.

[9] C. Giuliano, A. Lavelli, D. Pighin, and L. Romano, "FBK-IRST: Kernel methods for semantic relation extraction," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 141–144.

[10] M. d'Aquin, M. Sabou, E. Motta, S. Angeletou, L. Gridinoc, V. Lopez, and F. Zablith, "What can be done with the semantic web? an overview watson-based applications." in *Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives (SWAP2008)*, ser. CEUR Workshop Proceedings, A. Gangemi, J. Keizer, V. Presutti, and H. Stoermer, Eds., vol. 426. Rome, Italy: CEUR-WS.org, December 15-17 2008.

[11] J. Lehmann, J. Schppel, and S. Auer, "Discovering unknown connections - the dbpedia relationship finder," in *CSSW*, ser. LNI, S. Auer, C. Bizer, C. Mller, and A. V. Zhdanova, Eds., vol. 113. GI, 2007, pp. 99–110.

[12] M. d'Aquin, E. Motta, M. Sabou, S. Angeletou, L. Gridinoc, V. Lopez, and D. Guidi, "Toward a new generation of semantic web applications," *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 20–28, 2008.

[13] A. Weichselbraun, G. Wohlgenannt, A. Scharl, M. Granitzer, T. Neidhart, and A. Juffinger, "Discovery and evaluation of non-taxonomic relations in domain ontologies," *International Journal of Metadata, Semantics and Ontologies*, forthcoming.