A Resource-Poor Approach for Linking Ontology Classes to Wikipedia Articles

Nils Reiter
Matthias Hartung
Anette Frank
University of Heidelberg (Germany)
email: `reiter@cl.uni-heidelberg.de`

Abstract

The applicability of ontologies for natural language processing depends on the ability to link ontological concepts and relations to their realisations in texts. We present a general, resource-poor account to create such a linking automatically by extracting Wikipedia articles corresponding to ontology classes. We evaluate our approach in an experiment with the Music Ontology. We consider linking as a promising starting point for subsequent steps of information extraction.

1 Introduction

Ontologies are becoming increasingly popular as a means for formal, machine-readable modelling of domain knowledge, in terms of concepts and relations. Linking ontological concepts and relations to their natural language equivalents is of utmost importance for ontology-based applications in natural language processing. Providing larger quantities of text that clearly belongs to a given ontological concept is a prerequisite for further steps towards ontology population with relations and instances. We thus consider this work as a point of departure for future work on populating and lexicalizing ontologies, and their use in semantic processing.

In this paper we present a method that provides relevant textual sources for a domain ontology by linking ontological classes to the most appropriate Wikipedia articles describing the respective ontological class. The paper is structured as follows: We discuss related work in Section 2. Section 3 presents our method for linking ontology classes to Wikipedia articles. The method is implemented and tested using the music ontology (Raimond et al., 2007) and a Wikipedia dump of 2007. We present this experiment in Section 4 and its evaluation in Section 5. Section 6 concludes and gives an outlook on directions of future work.

2 Related Work

Our goal is to detect the most appropriate Wikipedia article for a given ontology class. As Wikipedia is a domain-independent resource, it usually contains many more senses for one concept name than does a domain-specific ontology. Thus, one of the challenges we meet is the need for disambiguation between multiple candidate articles with respect to one specific ontology class.[1] Therefore, we compare our approach to previous work on sense disambiguation. Since in our approach, we aim at minimizing the degree of language- and resource dependency, our focus is on the amount of external knowledge used.

One method towards sense disambiguation that has been studied is to use different kinds of text overlap: Ruiz-Casado et al. (2005) calculate vector similarity between a Wikipedia article and WordNet glosses based on term frequencies. Obviously, such glosses are not available for all languages, domains and applications. Wu and Weld (2007) and Cucerzan (2007) calculate the overlap between contexts of named entities and candidate articles from Wikipedia, using overlap ratios or similarity scores in a vector space model, respectively. Both approaches disambiguate named entities using textual context. Since our aim is to *acquire* concept-related text sources, these methods are not applicable.

A general corpus-based approach has been proposed by Reiter and Buitelaar (2008): Using a domain corpus and a domain-independent reference corpus, they select the article with the highest domain relevance score among multiple candidates. This approach works reasonably well but relies on the availability of domain-specific corpora and fails at selecting the appropriate among multiple in-domain senses. In contrast, our resource-poor approach does not rely on additional textual resources, as ontologies usually do not contain contexts for classes.

[1] Mihalcea (2007) shows that Wikipedia can indeed be used as a sense inventory for sense disambiguation.

3 Linking Ontology classes to Wikipedia articles

This section briefly reviews relevant information about Wikipedia and describes our method for linking ontology classes to Wikipedia articles. Our algorithm consists of two steps: (i) extracting candidate articles from Wikipedia and (ii) selecting the most appropriate one. The algorithm is independent of the choice of a specific ontology.[2]

3.1 Wikipedia

The online encyclopedia Wikipedia currently comprises more than 2,382,000 articles in about 250 languages. Wikipedia is interesting for our approach because it is semi-structured and articles usually talk about one specific topic.

The structural elements in Wikipedia that we rely on are links between articles, inter-language links, disambiguation and redirect pages. Inter-language links refer to an article about the same topic in a different language. Disambiguation pages collect the different senses of a given term. Redirect pages point to other pages, allowing for spelling variations, abbreviations and synonyms.

3.2 Extracting the candidate articles

The first step of our algorithm is to extract candidate articles for ontology classes. The method we employ is based on Reiter and Buitelaar (2008). The algorithm starts with the English label L_C of an ontology class C, and tries to retrieve the article that bears the same title.[3] Any Wikipedia page P retrieved by this approach falls into one of three categories:

1. P is an *article*: The template `{{otheruses}}` in the article indicates that a disambiguation page exists which lists further candidate articles for C. The disambiguation page is then retrieved and we proceed with step 2. Otherwise, P is considered to be the only article for C.

2. P is a *disambiguation* page: The algorithm extracts all links on P and considers every linked page as a candidate article.[4]

3. P is a *redirect* page: The redirect is being followed and the algorithm checks the different cases once again.

3.3 Features for the classifier

We now discuss the features we apply to disambiguate candidate articles retrieved by our candidate extraction method with regard to the respective ontology class. Some features use structural properties of both Wikipedia and the ontology, others are based on shallow linguistic processing.

[2] It is still dependent on the language used for coding ontological concepts (here English). In future work we aim at bridging between languages using Wikipedia's inter-language links or other multi-lingual resources.

[3] We use common heuristics to cope with CamelCase, underscore whitespace alternation etc.

[4] Note that, apart from pointing to different readings of a term, disambiguation pages sometimes include pages that are clearly not a sense of the given term. Distinguishing these from true/appropriate readings of the term is not trivial.

Domain relevance

Wikipedia articles can be classified according to their domain relevance by computing the proportion of domain terms they contain. In this paper, we explore several variants of matching a set of domain terms against the article in question:

Class labels. The labels of all concepts in the ontology are used as a set of domain terms.

- We extract the nouns from the POS-tagged candidate article. The relative frequency of domain terms is then computed for the complete article and for nouns only, both for types and for tokens.
- We compute the frequency of domain terms in the first paragraph only, assuming it contains domain relevant key terms.
- The redirects pointing to the article in question, i.e., spelling variations and synonyms, are extracted. We then compute their relative frequency in the set of class labels.

Comments. As most ontologies contain natural language comments for classes, we use them to retrieve domain terms. All class comments extracted from the ontology are POS-tagged. We use all nouns as domain terms and compute their relative frequencies in the article.

Class vs. Instance

It is intuitively clear that a class in the ontology needs to be linked to a Wikipedia article representing a class rather than an instance.[5] We extract the following features in order to detect whether an article represents a class or an instance, thus being able to reject certain articles as inappropriate link targets for a particular class.

Translation distance. Instances in Wikipedia are usually named entities (NEs). Thus, the distinction between concepts and instances can, to a great extent, be rephrased as the problem of NE detection. As our intention is to develop a linking algorithm which is, in principle, language-independent, we decided to rely on the inter-language links provided by Wikipedia. The basic idea is that NEs are very similar across different languages (at least in languages using the same script), while concepts show a greater variation in their surface forms across different languages. Thus, for the inter-language links on the article in question that use latin script, we compute the average string similarity in terms of Levenshtein Distance (Levenshtein, 1966) between the title of the page and its translations.

Templates. Wikipedia offers a number of structural elements that might be useful in order to distinguish instances from concepts. In particular, the `infobox` template is used to express structured information about instances of a certain type and some of their properties. Thus, we consider articles containing an `infobox` template to correspond to an instance.

[5] We are aware of the fact that the distinction between classes and instances is problematic on both sides: Ontologies described in OWL Full or RDF do not distinguish clearly between classes and instances and Wikipedia does not provide an explicit distinction either.

4 Experiment

4.1 The Music Ontology

We test our approach on the Music Ontology (MO) (Raimond et al., 2007). The MO has been developed for the annotation of musical entities on the web and provides capabilities to encode data about artists, their albums, tracks on albums and the process of creating musical items.

The ontology defines 53 classes and 129 musical properties (e.g. melody) in its namespace, 78 external classes are referenced. Most of the classes are annotated with comments in natural language. The MO is connected to several other ontologies (W3C time[6], timeline[7], event[8], FOAF[9]), making it an interesting resource for domain relevant IE tasks and generalisation of the presented techniques to further domains. The MO is defined in RDF and freely available[10].

4.2 Experimental Setup

The experiment is divided into two steps: candidate page selection and classification (see Section 3). For candidate selection we extract Wikipedia pages with titles that are near-string identical to the 53 class labels. 28 of them are disambiguation pages. From these pages, we extract the links and use them as candidates. The remaining 25 are directly linked to a single Wikipedia article.

To test our classification features, we divide the overall set of ontology classes in training and test sets of 43 and 10 classes, respectively, that need to be associated with their most appropriate candidate article. We restrict the linking to one most appropriate article. For the classification step, we extract the features discussed in Section 3.

Since the candidate set of pages shows a heavily skewed distribution in favour of negative instances, we generate an additional training set by random oversampling (Batista et al., 2004) in order to yield training data with a more uniform distribution of positive and negative instances.

5 Evaluation

For evaluation, the ambiguous concepts in the ontology have been manually linked to Wikipedia articles. The linking was carried out independently by three annotators, all of them computational linguists. Each annotator was presented the class label, its comment as provided by the ontology and the super class from which the class inherits. On the Wikipedia side, all pages found by our candidate extraction method were presented to the annotators.

The inter-annotator agreement is $\kappa = 0.68$ (Fleiss' Kappa). For eight concepts, all three annotators agreed that none of the candidate articles is appropriate and for ten all three agreed on the same article. These figures underline the difficulty of the problem, as the information contained in domain ontologies and Wikipedia varies substantially with respect to granularity and structure.

[6] `www.w3.org/TR/owl-time/`

[7] `motools.sourceforge.net/timeline/timeline.html`

[8] `motools.sourceforge.net/event/event.html`

[9] `xmlns.com/foaf/spec/`

[10] `musicontology.com`

Candidate article selection. Candidate selection yields 16 candidate articles per concept on average. These articles contain 1567 tokens on average. The minimal and maximal number of articles per concepts are 3 and 38, respectively.

Candidate article classification. We train a decision tree[11] using both the original and the oversampled training sets as explained above.

Table 1: Results after training on original and over-sampled data

	Positives		**Negatives**		**Average**	
	orig.	samp.	orig.	samp.	orig.	samp.
P	1	0.63	0.87	0.97	0.94	0.80
R	0.17	0.83	1	0.91	0.58	0.87
F	0.27	0.71	0.93	0.94	0.75	0.83

Table 1 displays precision, recall and f-score results for positive and negative instances as well as their average. As the data shows, oversampling can increase the performance considerably. We suspect this to be caused not only by the larger training set, but primarily by the more uniform distribution.

The table shows further that the negative instances can be classified reliably using the original or oversampled data set. However, as we intend to select positive appropriate Wikipedia articles rather than to deselect inappropriate ones, we are particularly interested in good performance for the positive instances. We observe that this approach identifies positive instances (i.e., appropriate Wikipedia articles) with a reasonable performance when using the oversampled training set. It is noteworthy that not a single feature performs better than with an f-measure of 0.6 *when used alone*. The figures shown in Table 1 are obtained using the combination of all features.

Table 2: Results for combination of best features only

	Positives	**Negatives**
P	0.60	1.00
R	1.00	0.88
F	0.75	0.94

In Table 2, we present the results for the best performing features taken together (using oversampling on the training set): `nountypes-classlabels` (F-measure: 0.6), `langlinks` (0.5), `redirects-classlabels` (0.5), `nountokens-classlabels` (0.44), `fulltextclasslabels` (0.44). Recall improves considerably, while there is a small decrease in precision.

6 Conclusions

We have presented ongoing research on linking ontology classes to appropriate Wikipedia articles. We consider this task a necessary step towards automatic ontology lexicalization and population from texts.

[11] We used the ADTree implementation in the Weka toolkit `www.cs.waikato.ac.nz/ml/weka/`.

The crucial challenge in this task is to deal with the high degree of ambiguity that is introduced by the fact that Wikipedia covers a large amount of fine-grained information for numerous domains. This leads to a great number of potential candidate articles for a given ontology class.

Our approach to this problem is independent of the particular ontology that is used as a starting point. Moreover, it merely depends on a set of rather shallow but effective features which can be easily extracted from the domain ontology and Wikipedia, respectively. From the results we derived in our experiments with the Music Ontology, we conclude that our approach is feasible and yields reasonable results even for small domain ontologies, provided we can overcome highly skewed distributions of the training examples due to an overwhelming majority of negative instances. In future work we will apply the methods described here to different domain ontologies and use the selected Wikipedia articles as a starting point for extracting instances, relations and attributes.

Acknowledgements. We kindly thank our annotators for their effort and Rüdiger Wolf for technical support.

References

Batista, G., R. Prati, and M. C. Monard (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations 6*, 20–29.

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proc. of EMNLP*, Prague.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*, 707–710.

Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. In *Proc. of NAACL-07*, Rochester, New York, pp. 196–203.

Raimond, Y., S. Abdallah, M. Sandler, and F. Giasson (2007). The Music Ontology. In *Proc. of the 8th International Conference on Music Information Retrieval*, Vienna, Austria.

Reiter, N. and P. Buitelaar (2008). Lexical Enrichment of Biomedical Ontologies. In *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration*. IGI Global, *to appear*.

Ruiz-Casado, M., E. Alfonseca, and P. Castells (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Proc. of the 3rd Atlantic Web Intelligence Conference*, Volume 3528, Lodz, Poland, pp. 380–385.

Wu, F. and D. S. Weld (2007). Autonomously Semantifying Wikipedia. In *Proc. of the Conference on Information and Knowledge Management*, Lisboa, Portugal.