Three steps for creating high-quality ontology lexica

John McCrae, Philipp Cimiano

Cognitive Interaction Technology Center of Excellence, University of Bielefeld
Universitätsstraße, D-33615 Bielefeld, Germany
{jmccrae,cimiano}@cit-ec.uni-bielefeld.de

Abstract

Sophisticated NLP applications working on particular domains require rich information on both the linguistic properties of words and the semantics of these words. We propose a three-step methodology for the creation of high-quality ontology-lexica which combine detailed syntactic information with deep semantic information about words and their associated meanings. Our proposed methods consists of three steps: first we rely on a standard NLP pipeline to create a preliminary version of the ontology lexicon automatically. In this step, the automatically created lexicon is linked to existing legacy lexical resources. The second step involves referencing existing lexical and semantic resources and importing data. Finally, a manual review step is required that is supported by a novel editor to facilitate the inspection and manual validation and modification and thus continuous refinement and improvement of the ontology lexicon.

1. Introduction

For many sophisticated NLP applications, such as question answering (Unger and Cimiano, 2011), natural language generation and machine translation (Beale et al., 1995; McCrae et al., 2011a), which work with text in specific domains, the creation of domain-specific lexical resources. However, the process of creating such resources often involves a significant amount of manual effort. In this paper, we propose a three-step method for creating *ontology-lexica* (Cimiano et al., 2007; Peters et al., 2007). Ontology lexica essentially specify how words, phrases etc. should be interpreted in the context of a given domain ontology and are thus crucial for ontology-based NLP applications. In particular, we propose the creation of ontology lexica by firstly creating an initial resource using a fully automatic process that builds in part on statistical natural language processing techniques for aspects such as identification of part-of-speech. Secondly, the process refers to existing resources and includes extra information from these sources in a semi-automatic manner, consulting the user primarily when ambiguity exists. Finally, the process involves a manual review of the results, allowing the user to correct errors that may have been introduced by the automatic tools. In such a way we envision that a resource such as an ontology-lexicon can be created quickly and easily for specific domains. We present this work in reference to a system called *lemon source* that allows for the creation of ontology-lexica using the *lemon* (McCrae et al., 2012a) ontology-lexicon format.

2. Ontology-lexicon models

Ontologies are widely used to represent semantics and the OWL format (McGuinness et al., 2004) has provided a standard format that has lead to the creation of a large number of ontological resources on the web, creating the *Semantic Web*. These ontologies have been applied to a number of natural language processing tasks. However, as noted by Buitelaar et al. (2009), the linguistic information contained in ontologies is typically not sufficient for NLP applications. Thus, in the past we have proposed the *lemon* model for formalizing and representing lexica which enrich ontologies with information about how the ontology elements are realized in different natural languages. *lemon* builds on existing work on Semantic Web lexical resources, in particular the LexInfo (Cimiano et al., 2011) and the LIR models (Montiel-Ponsoda et al., 2008) as well as the lexicon modelling framework, LMF (Francopoulo et al., 2006). The model thus aims to provide a richer description of lexico-linguistic information related to classes, properties and individuals in the ontology. In particular the *lemon* model contains a core set of elements describing a path between the ontological entity and the (string) label ("core path") consisting of **lexical entries** uniquely identified by URIs and available on the web as RDF data (Lassila et al., 1998). Lexical entries themselves consist of **lexical forms**, which record the inflectional variants of an entry and **lexical senses**, consisting of a **reference** to the ontology. A lexical form may have multiple **representations** in different scripts and/or orthographies and may have different phonetic representations and a lexical sense may be further described by pragmatic constraints.

In addition, the *lemon* model has a number of modules that extend the core path to handle the linguistic data required by these applications, in particular the following modules are used:

- **Linguistic Description**: Allowing for elements to be assigned to linguistic categories, e.g., of gender, case, number.
- **Variation**: Allows elements within the ontology-lexicon to be linked to elements of the same or other ontology-lexica.
- **Phrase structure**: Description of the decomposition of terms into other terms.
- **Syntax and Argument Mapping**: Consisting of syntactic frames and their correspondence to semantic predicates in the ontology.
- **Morphology**: Compact representation of form variants for highly synthetic languages.

Further details are described in the *lemon* cookbook[1].

3. Methodology

We propose a methodology for the creation of ontology-lexica as a three-step process illustrate each of the steps by an example.

3.1. Automatic resource creation

The first step we apply is to use automatic tools to create a skeleton resource that we can further enrich at the later stage. The methodology for this was described in (McCrae et al., 2011b) and we will recap it here. As input we assume that we have a resource that contains a set of terms for the domain, such as the labels for an ontology in OWL[2]. As an example, currently our system applies the following sub-steps:

- **Tokenization of multi-word terms**: This step involves analysing the label to see if it consists of multiple words. For European languages this is achievable with simple finite state automata, but it is often more complex for languages such as Chinese, Japanese or Korean (e.g., Wu and Fung (1994)).
- **Part-of-speech detection**: The next step is then to apply a part of speech tagger to deduce the part-of-speech of the word(s) in the label. In particular we use the Stanford Tagger (Toutanova and Manning, 2000).
- **Stemming**: We then normalize inflected forms of words in our label by means of a stemmer. For English we use the Stanford Tagger's stemmer and for other languages the Snowball stemmer[3].
- **Decompounding**: For some languages, notably German and Dutch, we need to break up compound words into their individual words, for example breaking "Qualitätsverbesserungskommission" ("quality improvement committee") into "Qualität", "Verbesserung" and "Komission." This is in practice achieved by applying the stemmer multiple times using the Viterbi algorithm.
- **Parsing**: After this we apply a parser to deduce the structure of a phrase if it consists of multiple words. In particular we use the Stanford Parser (Klein and Manning, 2003).
- **Frame detection**: We also detect for a verb or relational expression (such as "capital of") the number and kind of arguments it can take. Currently this is performed using rules based on the phrase structure/part-of-speech of the term, but could also be achieved by corpus analysis.
- **Subterm detection**: We search for common subterms across multiple input terms, extract these subterms from them and introduce new lexical entries for these subterms.
- **Term variation**: Here we use syntactic rules to suggest variants for terms; for example, we find in English that terms of the form $\texttt{NN}_1$ $\texttt{NN}_2$ can often also be expressed as $\texttt{NN}_2$ `of the` $\texttt{NN}_1$ such as "prostate cancer" and "cancer of the prostate."
- **Semantic relation induction**: We can use the lexical form of a word to induce semantic relationships between the terms. Currently, we induce hypernym relationships if two terms are subterms of one another, for example "personal profile document" is a type of "document."

Our system currently supports English and German, with partial support[4] for French, Dutch, Spanish and Chinese, which we plan to increase to full support.

3.2. Semi-automatic resource re-use

After having created a preliminary version of the ontology lexicon using automatic processing, we proceed to link this resource to other external resources. In particular we use two kinds of resources: machine-readable dictionaries, which have already been aligned to the *lemon* model (McCrae et al., 2012c) and semantic resources we find from the Web. More specifically, we use two lexical resources: WordNet (Fellbaum, 2010) and Wiktionary[5]. We rely on the following criteria to search for possibly aligned terms in the resources:

- The canonical (lemma) form is the same
- The part-of-speech is the same if present
- The two entries do not have contradictory values for a property, e.g., different grammatical genders
- The entries do not have a contradictory inflected form, e.g., a different plural form

It was found in (McCrae et al., 2012c) that 21.6% of WordNet entries could be matched to Wiktionary pages using this method of which 97.2% were unambiguous (in that there was only a single candidate Wiktionary page); the remaining WordNet entries has no equivalent in Wiktionary. This shows that the overlap between WordNet and Wiktionary is rather low in general.

The second kind of resource we attempt to link to are semantic resources, which we discover by using semantic web source engines such as the Watson search engine (d'Aquin et al., 2007). We detect similar concepts in these resources using a vector space model alignment algorithm similar to the one described by Trillo et al. (2007).

For both methods, we automatically link the ontology-lexicon to the relevant resource if it can be done so unambiguously. However, if multiple candidates are found by

[1] `http://lexinfo.net/lemon-cookbook.pdf`

[2] We note that there is significant effort required to identify the terminology required for a specific task, however automatic methods for extracting a term from a domain can be used.

[3] `http://snowball.tartarus.org`

[4] Generally, a trained model for the parser or tagger is not available

[5] `http://www.wiktionary.org/`

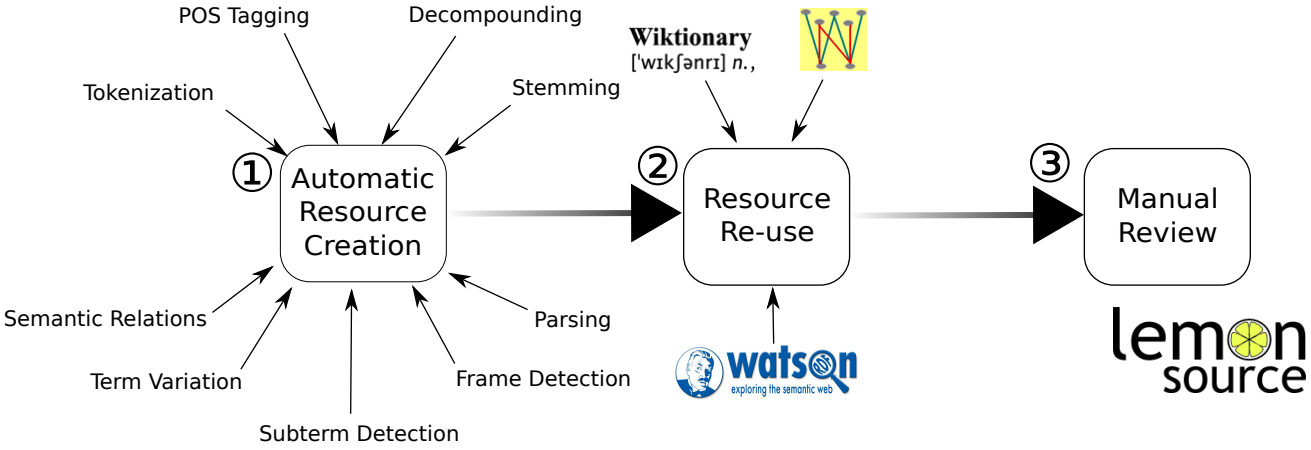

Figure 1: The three step procedure for creating an ontology-lexicon

the linking procedure we consult the user to allow them to select which resource should be used.

3.3. Manual review

The final step in the creation of an ontology-lexica is to manually review the result. While this could be achieved by examining the serialised (i.e., XML) form of the resulting resource, this would not be practical as many users, especially those with a non-technical background, would have trouble understanding the form of the resource, and it may be difficult to keep track of the progress. For this reason we have created a web application we call *lemon source* (McCrae et al., 2012b) that displays the result of the automatic and semi-automatic extraction and allows the user to edit the resulting entries in an intuitive manner. In addition, the system also allows the creation of meta-data about the ontology-lexicon, in particular assigning each entry to a number of statuses, such as "for review", "accepted" or "rejected." Lemon source allows users to collaboratively work on the lexicon by allowing their shared, remote use and by making updates made by one client immediately available to all clients, such as in Cunningham et al. (2003).

4. Discussion

This methodology for the creation of ontology-lexica is necessary for the sophisticated ontology-based NLP applications that we target, as we find that neither automatic methods, existing resources nor manual resource creation are sufficient to meet the challenge of creating high-quality lexica.

In the case of automatic ontology-lexicon induction, we would ideally hope that the result would be accurate and sufficient for the nature of the tasks we envision. However, while the systems we use have very high accuracy, they cannot said to be perfect. Our system achieves between 99.1% and 81.5% precision depending on the ontology (see (McCrae et al., 2011b)). A fundamental issue with the creation of a language resource by automatic methods is that any text processing system that uses an automatically generated language resource could achieve at least as good performance by directly integrating the tools used to create the language resource. As an example of this, consider that our text processing system needs to know the part-of-speech of terms used within a phrase; a language resource could extract this using a part-of-speech tagger, as we do. However, a statistical part-of-speech tagger will produce more information, such as the probability of individual words being tagged with a particular part-of-speech and other potential candidate taggings, which could be utilized by the end system. It is of course possible that we could include such information in the language resource at the risk of needlessly bloating the language resource in a manner that could make it difficult to use in practise. Nevertheless, materializing this information into an ontology-lexicon has the potential to reduce costs overall as people interested in exploiting an ontology for a given NLP application could download and reuse existing lexica instead of creating them from scratch.

In the case of reusing existing language resources, we have a clear advantage in that we can assume that these resources are of very high quality and much less likely to contain errors. However, there is a clear issue that for domain terminology it is highly unlikely that the resources contain all required entries. This proves to be significant when applying text processing applications that are dependent on language resources to new domain. However, much of the necessary information for these applications cannot easily be deduced by automatic methods, especially the extraction of specific relations between concepts and relationships involving multiple concepts (Zhou, 2007).

Finally, manual editing systems are ultimately necessary for the creation of high-quality resources. However, the creation of a complex language resource is extremely time-consuming and often requires users with specific training in linguistic resources. Moreover, it has been shown that complex annotation schemes like those required for structured resources like ontology-lexica lead to a lot of errors (Butler et al., 2000). As such, reducing the complexity of the scheme and the amount of the resource that needs to be created is a key goal of the manual annotation (Bayerl et al., 2003), and this can be carried out by incorporating automatic assistance (Smith et al., 2008).

5. Conclusion

We have proposed a three-step methodology for the creation of high-quality ontology lexica based on the use

of automatic tools, semi-automatic re-use of existing language resources and manual review, and presented a detailed overview of *lemon source*, an implemented web application that supports this methodology.. Each of these steps is extremely valuable for creating such resources, but the single steps have significant and complementary costs. Thus, by combining all three methodologies, high quality language resources which have high coverage and high accuracy for particular domains, can be quickly created.

6. References

P.S. Bayerl, H. Lüngen, U. Gut, and K.I. Paul. 2003. Methodology for reliable schema development and evaluation of manual annotations. In *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003)*.

S. Beale, S. Nirenburg, and K. Mahesh. 1995. Semantic analysis in the Mikrokosmos machine translation project. In *Proceedings of the 2nd Symposium on Natural Language Processing*, pages 297–307.

P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek. 2009. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, pages 111–125.

T. Butler, S. Fisher, G. Coulombe, P. Clements, I. Grundy, S. Brown, J. Wood, and R. Cameron. 2000. Can a team tag consistently?: Experiences on the Orlando project. *Markup Languages*, 2(2):111–125.

P. Cimiano, P. Haase, M. Herold, M. Mantel, and P. Buitelaar. 2007. LexOnto: A model for ontology lexicons for ontology-based NLP. In *Proceedings of the OntoLex07 Workshop at the 6th International Semantic Web Conference*.

P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. 2011. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

H. Cunningham, V. Tablan, K. Bontcheva, and M. Dimitrov. 2003. Language Engineering Tools for Collaborative Corpus Annotation. In *Proceedings of Corpus Linguistics 2003*, pages 80–87.

M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. 2007. Watson: a gateway for the semantic web. In *Proceedings of the 4th Annual European Semantic Web Conference*.

C. Fellbaum. 2010. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243.

G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the 2006 International Conference on Language Resource and Evaluation (LREC)*, pages 233–236.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

O. Lassila, R.R. Swick, et al. 1998. Resource description framework (RDF) model and syntax specification. Technical report, W3C Recommendation.

J. McCrae, M. Espinoza, Montiel-Ponsoda, G. Aguado-de Cea, and P. Cimiano. 2011a. Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Fifth workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5)*.

J. McCrae, D. Spohr, and P. Cimiano. 2011b. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC)*.

J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012a. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*.

J. McCrae, E. Montiel-Ponsoda, and P. Cimiano. 2012b. Collaborative semantic editing of linked data lexica. In *Proceedings of the 2012 International Conference on Language Resource and Evaluation (LREC)*.

J. McCrae, E. Montiel-Ponsoda, and P. Cimiano. 2012c. Integrating wordnet and wiktionary with lemon. In *Workshop on Linked Data in Linguistics 2012*.

D.L. McGuinness, F. Van Harmelen, et al. 2004. OWL web ontology language overview. Technical report, W3C recommendation.

E. Montiel-Ponsoda, G.A. de Cea, A. Gómez-Pérez, and W. Peters. 2008. Modelling multilinguality in ontologies. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING)*, pages 67–70.

W. Peters, E. Montiel-Ponsoda, G. Aguado de Cea, and A. Gómez-Pérez. 2007. Localizing ontologies in owl.

N. Smith, S. Hoffmann, and P. Rayson. 2008. Corpus tools and methods, today and tomorrow: Incorporating linguists manual annotations. *Literary and Linguistic Computing*, 23(2):163–180.

K. Toutanova and C.D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 conference on Empirical methods in natural language processing*, pages 63–70.

R. Trillo, J. Gracia, M. Espinoza, and E. Mena. 2007. Discovering the semantics of user keywords. *Journal of Universal Computer Science*, 13(12):1908–1935.

C. Unger and P. Cimiano. 2011. Pythia: Compositional meaning construction for ontology-based question answering on the semantic web. In *Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB 2011)*, pages 153–160.

D. Wu and P. Fung. 1994. Improving chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 180–181.

L. Zhou. 2007. Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3):241–252.