Multiple Instance Terminological Thesaurus with Central Management

Adam Rambousek

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
rambousek@fi.muni.cz

Abstract. This paper describes the design of the new specialized dictionary writing system for the creation and management of terminological thesaurus. To help with information sharing and terminology unification, the system also includes central node that keeps track of all the dictionary instances and synchronize data between them.

Keywords: terminology thesaurus; dictionary writing; DEB platform; Linked Data

1 Introduction

Specialists in any branch inevitably rely on domain-specific vocabulary as a basis for sharing exact terminology amongst professionals. Such detailed domain terminology cannot be included in general language dictionaries, which is why specialized terminology dictionaries are being built and managed. With the need to share information unambiguously in different languages, terminology dictionaries often link original terms to their translations. Taxonomic ordering of the terminology is described by means of term relations such as synonymy or hypernymy/hyponymy. In our new system, information about the terms is presented and visualized in a way that helps the readers (both specialists and the general public) to understand the meaning of the term and its usage in contexts.

As a pilot project, The Natural Language Processing Centre (NLP Centre) at the Faculty of Informatics, Masaryk University in cooperation with the Czech Office for Surveying, Mapping and Cadastre (CUZK) has developed a new system for building and extending a specialized terminology thesaurus for the domain of land surveying and land cadastre. The project consists of several tightly interconnected parts—a web-based application to create, edit, browse and visualize the terminology thesaurus, and a set of tools to build large corpora of domain oriented documents which allows for the detection of newly emerging terms, or terms missing from the thesaurus.

In the follow-up project, the developed application for creation and editing of terminology thesaurus will be updated to be generally usable for any domain. Thus any organization may re-use the same system for terminology dictionary.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 71–76, 2018. © Tribun EU 2018

However, with several applications running in daily use, same term may appear in various thesauri. Also, users might want to inter-link connected terms between dictionaries. To handle the management of thesaurus instances and links between them, new central management system is in development.

1.1 The DEB platform

Both the thesaurus application and the central management system are developed using the universal dictionary writing system developed at the NLP Centre (Faculty of Informatics, Masaryk University). The system is called Dictionary Editor and Browser, or the DEB platform [3,4]. Since 2005, the DEB platform was applied in more than 10 large international research projects. Large-scale applications based on the DEB platform include the lexicographic workstation for the development of the Czech Lexical Database [2] with detailed morpho-syntactic information on more than 213 thousands Czech words, or the complex lexical database Cornetto combining the Dutch wordnet, an ontology, and an elaborate lexicon [5]. Currently ongoing projects include Pattern Dictionary of English Verbs tightly interlinked with the corpus evidence [6], Family names in Britain and Ireland [1] providing detailed investigations for over 45,000 surnames to be published by Oxford University Press, or the dictionary of the Czech Sign Language[1] with an extensive use of video recordings to present the signs [7].

The DEB platform is based on the client-server architecture, which brings along a lot of benefits. All the dictionary and interlinked data are stored on a server and a considerable part of the functionality is also implemented on the server-side, consequently the client application can be very lightweight. This approach provides very good tools for editor team cooperation; data modifications are immediately seen by all involved users. The DEB server also provides authentication and authorization tools.

2 Central management system

2.1 Thesaurus management

Central node keeps track of all the installed instances of thesaurus system. After the installation, local administrator of the thesaurus system fills in the metadata. Following details are stored at the central node:

- thesaurus instance ID,
- URL to access the data,
- name of the organization running this thesaurus,
- administrative contact,
- domain and content description.

[1] `http://www.dictio.info`

Metadata are sent to the central node, where they are stored as unverified data. New thesaurus instance is now available for search and inter-linking, although with the warning about unverified status. After the central node administrator contacts local organization and checks the details, new thesaurus instance may be switched to verified.

After registration, central system will start with periodical downloading of the thesaurus content (term list and entry details). This data serve as a back-up copy of each thesaurus and also are used for term linking and cross-reference checks, as described below.

2.2 Inter-linking terms

When the user of thesaurus application is adding new term, the same term is also checked in all registered instances. Request is sent to the central system and all the downloaded back-ups are queried for the new term. Consequently, central node returns the list of all instances that contain the entry for the given term.

After consulting the list, user may decide to copy details of one of the existing term entries. In such case, local thesaurus instance will request term entry details directly from the remote instance. Entry details are copied to the new entry, together with the link to source term entry. See Figure 1 with the chart explaining the process.

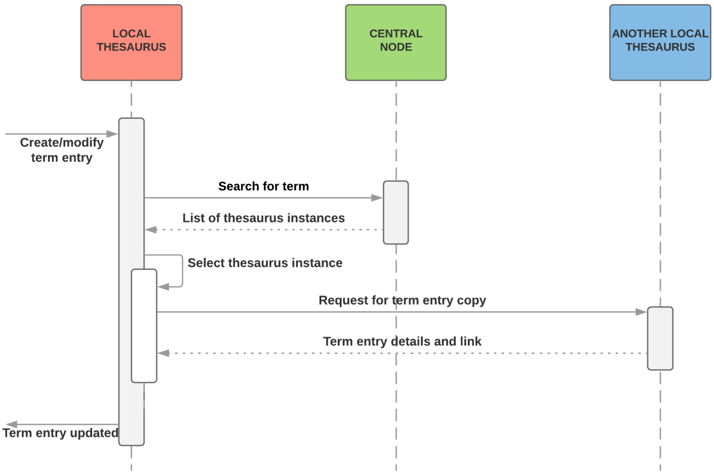

Fig. 1: Process of inter-linking thesaurus instances when creating new term entry.

2.3 Cross-reference checks

During the life cycle of various terminology thesaurus, existing entries are often updated, merged or split. These changes may also break the links between entries. For this reason, central management system periodically checks links between all the term entries, both in the same thesaurus instance, or in different instances.

If the central system detects updates in the link destination, editor-in-chief of the originating thesaurus instance is notified about the change. Consequently, editors will decide about the best action needed to keep term entries synchronized. See Figure 2 for the example of term entry containing link to external thesaurus instance.

```
<entry id="3611">
  <terms>
    <term lang="cs">stavba</term>
    <term lang="en">building</term>
  </terms>
  <refs>
    <ref type="external" system_id="https://terminologie.mvcr.cz"
      entry_id="895">stavba (Geoinfostrategie)</ref>
  </refs>
</entry>
```

Fig. 2: Term entry with link to entry in external thesaurus instance.

2.4 Official reference checks

Many terminology dictionaries are mentioned as the reference data in various official documents (laws, standards, regulations), or are derived from the official documents, e.g. term meaning is defined by the law. To support this kind of link, the thesaurus system provides special format of cross-reference links to official documents. Source data for the documents will be provided by the e-government office of the Ministry of the Interior and the cross-reference format was consulted to conform to future specification.

If the central system is notified by the external service that some official document was updated, all the entries in each thesaurus instance are checked. When an entry is found linking to the given document, editor-in-chief is notified and decides the best action to keep term entries in line with the official reference document. See Figure 3 for the example of term entry linking to the law where the term is defined.

```
<entry id="3761">
  <terms>
    <term lang="cs">stavební pozemek</term>
    <term lang="en">building site</term>
  </terms>
  <refs>
    <ref type="law" nr="183" year="2006" source="Sb">
      zákon č. 183/2006 Sb., o územním plánování a stavebním řádu (stavební zákon)
    </ref>
  </refs>
</entry>
```

Fig. 3: Term entry with link to the law where the term is defined.

2.5 Appearing new terms

As mentioned before, when the users create new term entry, they are provided with the list of thesaurus instances where the same term is existing. However, it may also happen that the same term appears in one of the thesaurus instances at later point.

To detect such case, central system is also periodically checking newly created term entries. If a new entry appears with the same term that is already existing, editors of all affected thesaurus instances are notified and asked to synchronize the term entries.

3 Conclusion

We have described enhancement of the lexicographic system for building and editing terminology thesaurus. The goal of the project currently in development is to inter-connect many thesaurus instances to the central management system. This organization will help to keep terminology synchronized between various domains and also in reference to the official government documents.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin OP VVV project CZ.02.1.01/0.0/0.0/16_013/0001781.

References

1. Hanks, P., Coates, R., McClure, P.: Methods for Studying the Origins and History of Family Names in Britain. In: Facts and Findings on Personal Names: Some European Examples. pp. 37–58. Acta Academiae Regiae Scientiarum Upsaliensis, Uppsala (2011)

2. Horák, A., Rambousek, A.: PRALED – A New Kind of Lexicographic Workstation. In: Przepiórkowski, A., Piasecki, M., Jassem, K., Fuglewicz, P. (eds.) Computational Linguistics: Applications, pp. 131–141. Springer (2013)
3. Horák, A., Rambousek, A.: DEB Platform Deployment – Current Applications. In: RASLAN 2007: Recent Advances in Slavonic Natural Language Processing. pp. 3–11. Masaryk University, Brno, Czech Republic (2007)
4. Horák, A., Rambousek, A.: Using DEB Services for Knowledge Representation within the KYOTO Project. In: Principles, Construction and Application of Multilingual WordNets, Proceedings of the Fifth Global WordNet Conference. pp. 165–170. Narosa Publishing House, New Delhi, India (2010)
5. Horák, A., Vossen, P., Rambousek, A.: A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In: Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing. pp. 1–15. Springer-Verlag, Haifa, Israel (2008)
6. Maarouf, I.E., Bradbury, J., Baisa, V., Hanks, P.: Disambiguating verbs by collocation: Corpus lexicography meets natural language processing. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
7. Rambousek, A., Horák, A.: Management and Publishing of Multimedia Dictionary of the Czech Sign Language. In: Biemann, C., Handschuh, S., Freitas, A., Meziane, F., Métais, E. (eds.) Natural Language Processing and Information Systems, NLDB 2015. pp. 399–403. Lecture Notes in Computer Science, Springer (2015). https://doi.org/10.1007/978-3-319-19581-0_37