Linking four heterogeneous language resources as linked data

Benjamin Siemoneit, John P. McCrae, Philipp Cimiano
Cognitive Interaction Technology, Center of Excellence, Bielefeld University
Bielefeld, Germany
bsiemone@techfak.uni-bielefeld.de
{jmccrae,cimiano}@cit-ec.uni-bielefeld.de

Abstract

The interest in publishing language resources as linked data is increasing, as clearly corroborated by the recent growth of the Linguistic Linked Data cloud. However, the actual value of data published as linked data is the fact that it is linked across datasets, supporting integration and discovery of data. As the manual creation of links between datasets is costly and therefore does not scale well, automatic linking approaches are of great importance to increase the quality and degree of linking of the Linguistic Linked Data cloud. In this paper we examine an automatic approach to link four different datasets to each other: two terminologies, the *European Migration Network (EMN)* glossary as well as the *Interactive Terminology for Europe* (IATE), BabelNet, and the Manually Annotated Subcorpus (MASC) of the American National Corpus. We describe our methodology, present some results on the quality of the links and summarize our experiences with this small linking exercise We will make sure that the resources are added to the linguistic linked data cloud.

1 Introduction

Linked data has recently become a popular approach to publishing language resources on the Web. It has been argued (Chiarcos et al., 2013) that the linked data approach applied to language resources has important advantages, most notably its ability to break the limitations of classical resource types and to foster integration of data by linking data across resources.

As the manual creation of links between datasets is costly and therefore does not scale well, automatic linking approaches are of great importance to increase the quality and degree of linking of the Linguistic Linked Data cloud. In this paper we describe the results of a small project attempting to link four datasets of different types (two terminologies, one lexico-conceptual resource and one corpus). As terminological resources, we have considered the Glossary of the European Migration Network (EMN)[1] as well as the Interactive Terminology for Europe (IATE) [2]. They are both represented using the *lemon* model (McCrae et al., 2012). As lexico-conceptual resource we rely on BabelNet (Navigli and Ponzetto, 2012), which has been previously migrated into Linked Data (Ehrmann et al., 2014). As corpus we use the Manually Annotated Subcorpus (MASC) of the American National Corpus (Ide et al., 2008), which contains disambiguated links to BabelNet.

We describe how the datasets have been migrated to RDF and describe our methodology for linking the datasets at the lexical entry level and present a sampled evaluation of the quality of the induced links. We first use a simple technique based on strict matching of the canonical form of lexical entries in different resources. By this we then link the EMN to both IATE and BabelNet. MASC has been previously linked to BabelNet and we included these links into our version of MASC.

The paper is structured as follows: in the next Section 2 we briefly describe the models that have been used to represented the data as Linked Data. Section 3 describes how the datasets have been converted into RDF. Section 4 describes our methodology for linking and presents a sampled evaluation of the quality of the automatically induced links.

[1] http://ec.europa.eu/dgs/home-affairs/what-we-do/networks/european_migration_network/glossary/index_a_en.htm

[2] http://iate.europa.eu/

Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), pages 59–63,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

2 Models

We used two models to represent the datasets presented in this paper. The terminologies and dictionaries have been represented in RDF using the *lemon* model (Lexicon Model for Ontologies) (McCrae et al., 2012), which has been designed to represent lexical information relative to ontologies and other semantic structures such as terminologies. For the MASC corpus we used the NLP Interchange Format (NIF) (Hellmann et al., 2013), a stand-off annotation format for the representation of annotations of text for NLP applications. We briefly describe these models in the following:

2.1 Lemon-OntoLex

The *lemon* (Lexicon Model for Ontologies) was proposed by McCrae et al. (McCrae et al., 2012) as a model for the representation of lexical information and has more recently been as a basis for the standardization work of the W3C Community Group on Ontology-Lexica.[3] The model revolves around the key concept of a *lexical entry*, which consists of a number of *forms* (e.g., 'plural form'), having different written or phonetic representations. The meaning of the lexical entry is specified by *reference* to some ontological concept. This relation is mediated by a *lexical sense*. In the case of the terminological resources EMN and IATE we model each terminological concept as a `skos:Concept` and model each term as a lexical entry that has the corresponding `skos:Concept` as *reference*.

2.2 NIF

Modelling corpus data such as MASC requires that we are capable of representing the annotations of this data in a compact and effective manner. The NLP Interchange Format (NIF) supports the annotation of text by using stand-off annotations represented as RDF. For this, it reifies strings in the document as RDF resources that refer to a specific character offset. For example, the URI `http://www.example.com/document.txt#char=3,7` would refer to the word occurring in the document which can be found at the path and server given in the URI, the fragment identifier follows RFC 5417 (Wilde, 2008) and identifies the word between the 3rd and 7th character. This annotation object can then be further annotated with properties from NIF such as the start and end index (to enable direct querying) or annotations from other schemas suitable for this corpus.

Resource	Size	Triples
IATE	8,081,142 terms	74,023,248
EMN	8,855 terms	106,283
MASC	506,768 words	8,650,723

Table 1: Size of the resources described in this paper without linking annotations.

3 Transformation to Linked Data

In this section we describe the transformation of the different datasets to RDF. The sizes of the resulting resources are given in Table 1 and are available for download at:

IATE `http://tbx2rdf.lider-project.eu/data/iate/`

EMN `http://data.lider-project.eu/emn/`

MASC `http://data.lider-project.eu/MASC-NIF/`

BabelNet `http://babelnet.org/rdf/`

The original data resources were primarily available as XML documents and thus conversion was for the most part the straightforward task of matching elements in an XML scheme to a appropriate RDF constructs. This was done by means of developing converters that parsed the XML and generated appropriate RDF. We will describe the details of the mapping in the next sections.

3.1 Transformation of EMN

The EMN glossary consists of 388 entries related to asylum and migration. Each entry is comprised of an English term with translations into 22 EU languages, a concept definition, semantic relations to other entries, explanatory comments and the source of the definition. We extracted the glossary from the HTML and converted it into linked data he *lemon* model.

A *lemon* `Lexicon` was created for each language. Then, for each EMN entry and for each of the available translations, a `LexicalEntry` was added to the respective `Lexicon`.

In *lemon*, `LexicalSense` objects are used for mapping terms to ontological entities. Although EMN entries are not RDF resources, we attached

[3] `http://www.w3.org/community/ontolex`

the URL of the respective entry as ontological reference to each sense.

The terms in EMN are not directly lemmas and so in order to incorporate them in a lexicon such as *lemon* we performed some preprocessing steps in order to obtain proper lexical entries: All additional information given in brackets or separated by special characters have been removed. The resulting strings were added as `LexicalForm` to their corresponding `LexicalEntry` objects.

3.2 Transformation of MASC

MASC contains 500K words of written and transcribed spoken language. Annotations for a variety of phenomena including BabelNet synset annotations are available in the Graph Annotation Format (GrAF) (Ide and Suderman, 2007). GrAF defines an XML serialization of graphs containing linguistic annotations. Graphs can, for example, be used to model the syntactic structure of the data, with nodes representing sentences, phrases etc. Leaf nodes refer to tokens in the primary data. Annotations can be attached to nodes and edges as feature structures.

In order to convert the corpus to NIF we first created a `nif:Context` for each primary data document. Nodes were then mapped to `nif:String` objects with normal RDF properties for a) a reference to the respective context object, b) the start and end indices of the chunk, c) the string representation of the chunk and d) all feature-value-pairs attached to the node.

3.3 Transformation of IATE

The IATE terminology is published using the TermBase Exchange (TBX, ISO 30042). We used the converter available under [4] to convert the IATE terminology into lemon-based RDF. As for EMN, terminological concepts were mapped to `skos:Concepts` and terms were mapped to lexical entries referring to the corresponding concept. In the IATE dataset, each concept has a *reliability code* and *subject field*, which were also represented as RDF. The language codes of terms were mapped to LexVo (de Melo, 2015) URIs.

4 Linking

4.1 Linking EMN to IATE

Concepts in the EMN datasets were linked to concepts in IATE by matching the written representation of the corresponding lexical entries in different languages. The number of languages for which the lexical entries for a given concept match was regarded as an indicator of the quality of the match, that is the more languages yield a match, the higher the quality of the induced link was expected to be.

In particular, EMN concepts were linked to IATE concepts by searching for string matches between corresponding EMN lexical entries and IATE lexical entries in multiple languages. In order to improve recall, we used Snowball stemming[5] for the eleven supported EU languages and transformed all strings to lowercase. The search was limited to IATE concepts associated with migration (subject field 2811).

Multiple IATE concepts can match a single EMN concept. In order to decide between candidate matches, we counted the number of languages for which each match holds and used this count as a measure for match plausibility (see Figure 1). We induced 3,028 links between EMN and IATE by considering all possible matches. Only considering the best match for each EMN concept resulted in 2,038 links (compare Table 2).

Resources	Number of links	Percentage of EMN	Precision
EMN-BabelNet	1,347	15%	69%
EMN-IATE (all matches)	3,082	35%	93%
EMN-IATE (best matches)	2,038	23%	94%

Table 2: Number of links between resources and precision of mapping.

4.2 Linking EMN to BabelNet

EMN concepts were linked to BabelNet by using the Babelfy (Moro et al., 2014) named entity linking service. Invoking the Babelfy disambiguation algorithm on the written representation of the lexical entries, we extracted all the synsets that Babelfy annotated the written representation with and considered only those annotations consisting of exactly one synset. A precision of 69% was determined by manually comparing concept definitions for a sample of 100 matches.

[4] `https://github/cimiano/tbx2rdf`

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

Resources	Number of links
IATE-EMN-BabelNet	700
EMN-BabelNet-MASC	37,405
IATE-EMN-BabelNet-MASC	7,794

Table 3: Number of transitive links added to resources.

On the basis of the existing linking between MASC and BabelNet and the above mentioned induced links between EMN and IATE (3,028, see Table 2) as well as between EMN and BabelNet (1,347, see Table 2), by transitive closure we were able to induce 700 links between IATE and BabelNet (via EMN as pivot), 37,405 links between EMN and MASC (via BabelNet as pivot) and 7,794 between IATE and MASC (via BabelNet and EMN as pivots). The results are summarized in Table 3.
To give an example, the EMN term 'visa' was linked to the matching term associated with IATE concept 3556819 and to BabelNet synset bn:00080087n, which in turn had been used to annotate 15 different tokens in MASC.

4.3 Linking precision

We evaluated the linking precision by manually evaluating a sample of 100 generated links. Precision of the linking is defined as the number of correctly created links divided by the number of generated links. Precision was determined by manually comparing terms, definitions and sources for a sample of matches: a link was judged as correct if the concepts share the same source or if their definitions don't contradict and there is no better matching concept.

The precision of the linking is shown in Table 2. The precision of linking EMN to IATE is quite high, which is due to the fact that they are terminologies and typically only contain one sense or meaning for a certain term / lexical entry. In contrast, BabelNet contains many possible senses for each lexical entry, so that the right sense among all the candidate senses needs to be found and this leads to errors.

We evaluated the precision of the induced links in dependence of the number of languages for which the written representations match. This analysis is shown in Figure 1. We observe that there is a clear improvement when considering links induced when the written representations for more than 5 languages match.

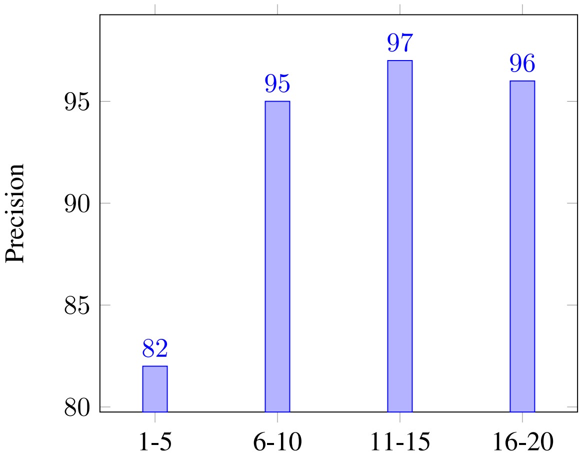

Figure 1: Precision of linking by number of languages matching for EMN-IATE mapping.

Finally, we evaluated the transitive linking and the results are presented in Table 3, we found that the two chains using one intermediate resource still maintained a large percentage of the links, as 52% of links from BabelNet to EMN could then be further extended to IATE. Furthermore, even using two intermediate resources still returned a useful number of links.

5 Conclusion

In this paper we have presented an experience report summarizing our experiences in developing an automatic approach to link four different language resources to each other. We have described a methodology that induces a link if the written representations of the lexical entries of the corresponding concepts match for a number of languages. We have shown that results are generally accurate, in particular when inducing links between terminologies. Further, the precision increases the more languages we require to have a match. Future work should be devoted to improving our methodology to increase both precision and recall of the generated links and thus reduce manual post-processing effort. Further, new methodologies for involving humans in the curation and validation of such links must be developed.

Acknowledgments

This work was funded under LIDER ("Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe"), an FP7 project refererence number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies. This work was also funded as part of the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University.

References

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. Towards open data for linguistics: Lexical linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

Gerard de Melo. 2015. Lexvo. org: Language-related information for the linguistic linked data cloud. *Semantic Web*, 6(4).

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John P. McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *In Proceedings of the Ninth International Conference on Language Resources and Evaluation*, volume 14, pages 401–408.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using linked data. In *Proceedings of the 12th International Semantic Web Conference*, pages 98–113.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Charles Fillmore. 2008. MASC: The manually annotated sub-corpus of American English. In *In Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

E. Wilde. 2008. URI fragment identifiers for text/plain media types. Technical report, Internet Engineering Task Force. RFC 5147.