Reusing Linguistic Resources: Tasks and Goals for a Linked Data Approach

Marieke van Erp

Abstract There is a need to share linguistic resources, but reuse is impaired by a number of constraints including lack of common formats, differences in conceptual notions, and unsystematic metadata. In this contribution, the five most important constraints and the tasks necessary to overcome these issues are detailed. These constraints lie in the design of linguistic resources, the way they are marked up and their metadata. These issues have also come up in a domain other than linguistics, namely in the semantic web, where the Linked Data approach proved useful. Experiences and lessons learnt from that domain are discussed in the light of standardisation and reconciliation of concepts and representations of linguistic annotations.

1 Introduction

Linguistic resources, which form the core of Natural Language Processing (NLP) research and applications, are expensive to generate. Most research uses some sort of manually annotated data which requires extensive human effort to create. The best corpora involve hundreds (if not thousands) of man-hours. Furthermore, even the results generated from the application of automated techniques still involve complex tuning and parametrisation of algorithms, making them resources not easily reproduced. Because of the expense in producing linguistic resources, the NLP community often shares and reuses its datasets. However, the reuse of linguistic resources is not straightforward and limited because of a number of constraints. In this contribution, these constraints are specified, their influence on reusability is explained and a path towards a solution is discussed that draws upon lessons learnt from the Link-

Marieke van Erp
Web and Media Group, Computer Sciences Department, VU University, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands e-mail: marieke@cs.vu.nl

ing Open Data Project.[1] At the time of writing this contribution, the LOD2[2] project publicly announced that they are developing an NLP Interchange Format (NIF)[3] to further interoperability between NLP tools, linguistic resources and annotations (Hellmann et al, this vol.). Interoperability is also addressed several contributions in this volume, including Chiarcos (this vol.), Eckart et al (this vol.), Windhouwer and Wright (this vol.), and it is one of the proclaimed goals of the Open Linguistics Working Group (Chiarcos et al, this vol., see there for related activities beyond the Semantic Web community). The emergence of such initiatives from the Linked Data community indicates that there is a desire from this community to use and reuse linguistic resources.
The following five constraints are discerned:

1. Linguistic resources are often designed for particular tasks (e.g., part-of-speech tagging, named entity recognition).
2. There are plethora of different mark-up languages, which are often not fully compatible between systems, much less between domains.
3. Each linguistic resource may use different conceptual models. For example, there are dozens of different part-of-speech tagsets (Petrov et al, 2011; Chiarcos, this vol.).
4. Existing linguistic resources often do not provide precise or machine readable definitions of the terminology they use, thus making it difficult to reuse them without manual investigation.
5. It is often difficult to obtain the full metadata around the creation of a resource. For example, metadata about the parameters set or the number of annotators used may be only documented within a paper or may be given at high level but not on per result basis.

Constraints 1 and 3 stem from choices made early on in the design of the resource. Constraint 2 is related to the design of the resource, but can often be overcome by mapping one annotation format to another. Constraints 4 and 5 pertain to the metadata associated with the resource.

1.1 Reuse Across Domains

Constraints 1 to 3 mostly bar combining resources and applying them to tasks that they have originally not been created for. Naturally, there is a trade-off between use and reuse, in particular in very domain-specific resources; the more specific a resource, the less reusable to others. But this does not mean that the linguistic community should not strive to facilitate reuse. The simplest scenario for which one wants to foster reuse is the case where one wants to combine two corpora that

[1] `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData` Retrieved: 30 November 2011

[2] `http://lod2.eu/` Retrieved: 30 November 2011

[3] `http://nlp2rdf.org` Retrieved: 30 November 2011

have been annotated for named entity recognition to create a larger training set for a statistical named entity recognition tool. Currently, the different conceptual models and mark-ups of corpora are not compatible with each other making this a difficult task, if possible at all.[4] In another scenario, one could use a resource that was for example developed for named entity recognition to train a part of speech tagger (as such resources often also contain part-of-speech information). As there is a particular annotation format used in the resource, which will rarely be the same as the annotation format that a tool supports, some form of data conversion is needed.

1.2 Reuse Across Communities

Constraints 4 and 5 bar reuse of resources by external parties as they make it more difficult for them to assess the data model, provenance of the data and quality of the data. This is a particularly discouraging issue for researchers or users who are not from the NLP community, but who would like to reuse linguistic resources for their applications. Because they are not familiar with the corpora, the threshold to start using them is high. This is mitigated by the fact that it is difficult for outsiders to assess the quality of the data, in particular with data that is generated (semi-)automatically.

Adopting a Linked Data approach can overcome these constraints. In addition to this, a Linked Data approach for linguistic resources may enrich existing annotations and create new opportunities for NLP tools that benefit from background knowledge.

In the remainder of this paper, an overview of how Linked Data approaches can be applied to linguistic resource reuse is given. This is done by going through the tasks necessary to convert two named entity recognition resources to Linked Data. But first, a brief overview of Linked Data is given in the next section.

2 Linked Data

The term ‘Linked Data’ is used to both describe the set of best practices for publishing and connecting structured data on the Web and to refer to the collection of data sets that have been published in this way so far. It gained traction with the Linked Open Data community project which promotes the publication of open data sets as RDF on the Web and by specifying links between instances of the different data sources. At the core of Linked Data lie the following four rules (Bizer et al, 2009a):

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.

[4] I mean possible without re-annotating the entire corpus.

3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

The number of data sets that is published has grown explosively over the past four years, and with it, use cases and applications have started to become available.

3 Tasks

In this section, the different tasks necessary to convert linguistic resource data to Linked Data are detailed. As a running example, the CoNLL 2003 language-independent named entity shared task (Tjong Kim Sang and Meulder, 2003) and the ACE 2005 entity detection and recognition task (Consortium, 2005) are used. The task of named entity recognition is taken as an example because it is a fairly well-understood and successful task in NLP. But even for this task, there are many differences in various resources. In the CoNLL shared task, the following four named entity types are discerned: PERSON, LOCATION,ORGANISATION, MISCELLANEOUS.

In the ACE 2005 annotation guidelines, the following seven entity types are discerned: FACILITY, GEO-POLITICAL ENTITY, LOCATION, ORGANISATION, PERSON, VEHICLE and WEAPON. In ACE, each of the seven main entity types also has several subtypes, for example *Facility* has subtypes *Airport*, *Building-Grounds*, *Path*, *Plant*, and *Subarea-Facility*.

Furthermore, in the CoNLL data set, the data representation is quite simple to understand and work with; each token is represented on a single line, followed by its part-of-speech tag and whether it is a named entity or not in IOB formatting. The ACE data is represented in XML, and contains, besides part-of-speech tag and entity information, also information about the type of reference the entity makes to something in the world and whether there are co-referring mentions. For reasons of space, only the main entity types and their mark-up are discussed.

3.1 Representation of Linguistic Annotations in RDF

Representing linguistic annotations in RDF is mostly a straight-forward transformation, and it is the least difficult of the conversion process for reuse. This step is successfully carried out if all information contained in the original format is represented in RDF. Linked Data is represented as an Subject-Predicate-Object structure, where the objects and subjects are either resources (e.g., objects that can be grounded in some ontology) or literals (e.g., simple strings or integers). It is very well possible to represent an annotated linguistic resource such as the CoNLL data set in RDF. First, one needs to create an ontology that specifies for each concept what it denotes. For the CoNLL example, our ontology would contain concepts such as words and sentences, the part of speech tagset, the chunk tagset and the named entity types. Then every sentence in the corpus can be represented as an RDF triple that contains a sentence-ID and the sentence text, for example we can create a triple `<conll:sent1> <rdfs:label> 'U.N. official Ekeus heads for Baghdad.'`. We then represent each word in the sentence as a set of RDF triples that contain the word, its position and links it to the sentence triple. We can then add triples for the part-of-speech tags that link each tag to each word, translate the chunk tags into RDF triples that tell us for each sentence at which position in the sentence a chunk starts and where it ends and likewise for each named entity where every entity mention starts and ends.

For the ACE 2005 data set, roughly the same process can be followed, although there the annotation already more extensive, it for example already encodes sentence IDs and the position of a word by character offset in the sentence. This can all be encoded in RDF triples.

When an RDF representation is created for every major element of the lexical resource, this first step in making linguistic resources more reusable has been achieved. The main requirement here is that a data format is used that is standardised and RDF is but one option. However, using RDF enables one to utilise semantic web technologies, which aid in dealing with the following task.

3.2 Mapping Annotations

The fact that two or more resources share the same format, does not automatically mean that they are integrated. Most resources will not be annotated following the exact same conceptual model, hence it is necessary to create mappings between the conceptual models of the resources as to have a unified annotation model. This task is carried out successfully when for each of the concepts in the one resource, it is clear what its counterpart is in the other resource. Much can be learnt from the ontology mapping work done in the Semantic Web (e.g., Euzenat et al 2011).

As mentioned at the beginning of this section, there is a discrepancy between the number of entity types that is discerned in the ACE 2005 and CoNLL tasks. Part of this can be explained that the CoNLL shared task explicitly addresses '**named** entities', and the ACE task only specifies 'entities', but the line between them is blurry. Indeed an entity type weapon may not always denote a named entity, but 'AK-47' can possibly be recognised as a name. One of the ACE facility subtypes is 'airports', in the CoNLL annotation, airports such as "Zaventem" are treated as locations. Another interesting issue is that of geo-political entity in ACE05, as this denotes terms such as "London" and "France" which can be locations, but they can also be used to refer to particular organisations (e.g., ***France*** *will combat market speculation*) or groups of persons (e.g., ***London*** *cleans up from riots*). In the CoNLL annotation, these would all be considered locations. In both the linguistic and Linked Data community, a satisfying high-level representation for such cases has not been found yet (cf. Recasens et al 2011 and Halpin et al 2010 respectively).

On a more fine-grained level, one also finds differences in entity annotations between corpora. In some corpora one will find for example that salutations, such as *Mr* and *Mrs*, are annotated as part of a person entity, in ACE and CoNLL they do not belong to the entity.

In general, it is easier to map more specific annotations such as those of ACE, to more general annotations, such as those of CoNLL, although this does mean that one cannot make use of the specificity of the ACE annotations anymore. In order to map the other way round, more information is needed, which is oftentimes not available in the more coarse-grained resource. There is no solution to this problem as yet. Possible solutions to this may be found in bootstrapping from the more specific resource to try to obtain finer granularity in the annotations in the lesser specific resources.

In order for interoperability and for a like-minded representation of named entities as Linked Data, it is important that such differences are either reconciled or mapped.

3.3 Grounding Annotations in Linked Data

When mapping annotation schemes and representing information as RDF, one has successfully made the transition to more reusability. However, Linked Data promises a large interlinked cloud of information. To ensure that linguistic resources find their way to other users and to enable tight integration with other resources, it is necessary to not only create mappings among linguistic resources but to also create mappings to other resources. For named entities one can very well imagine that mappings between entity types and DBpedia (Bizer et al, 2009b) are created and for part-of-speech tags that mappings are created that link to WordNet (Fellbaum, 1998).

For named entity recognition, one can also imagine grounding the entities in the LOD cloud,[5] for example by mapping locations to GeoNames,[6] persons to EntityPedia[7], organisations to WikiCompany[8] or general entities to DBpedia[9]. Numerous implementations for the task of entity linking are available (e.g., Hellmann et al, this vol.), but in general, this is not a trivial task and the tradeoff between costs and linking quality should be carefully investigated.

[5] `http://richard.cyganiak.de/2007/10/lod/` Retrieved: 30 November 2011

[6] `http://www.geonames.org` Retrieved: 30 November 2011

[7] `http://entitypedia.org/` Retrieved: 30 November 2011

[8] `http://wikicompany.org/` Retrieved: 30 November 2011

[9] `http://dbpedia.org/` Retrieved: 30 November 2011

3.4 Definition of Linguistic Resource Metadata

Besides converting the content of linguistic resources to RDF, it is also important to convert the meta-data about the resource to a machine-readable format. At the minimum level, this meta-data describes the conceptual model of the resources, i.e., describing the different elements of the resource. Ideally, this also includes information about how the data was collected, when it was annotated, and mappings to previous versions. As more (semi-)automatically annotated resources are being shared, quality assessment becomes more important. Before one decides to reuse a particular resource, it is good to know what the quality of this resource exactly is, so one can take this into account when working with it, preferably on instance level (e.g., one can add an RDF triple for each automatically generated annotation that indicates the confidence the annotation system has in the classification).

4 Related Work

Many initiatives have been undertaken to facilitate reuse of linguistic resources. Most of these have focused on creating standards for annotations (e.g., Pustejovsky et al, 2010) or making different annotation schemes interoperable, such as the mapping of different part-of-speech tagsets (Teufel, 1997; Petrov et al, 2011), and the development of RDF-based interchange formats, which was already mentioned in Section 1. Such an interchange format should facilitate exchange of annotations between different NLP tools. As currently the draft of the first version of the NLP Interchange Format is produced and the project is still underway, its impact on the NLP community and its resources is yet to be awaited.

A nice example of the added benefit of using semantic web technology for Natural Language Processing tasks is Mika et al (2008), who create a mapping between the CoNLL NER tagset and Wikipedia, and subsequently use this to improve the NER process.

5 Conclusion

In this contribution, the tasks and goals necessary to make linguistic resources reusable using semantic web technologies have been outlined. The main issues preventing reuse are standardisation and reconciliation of conceptual models. Solutions to these issues are sought in 1) converting existing annotation formats to RDF, 2) mapping annotations to each other and/or to a universal annotation scheme, 3) grounding the resources in the linked open data cloud, and 4) defining the metadata. These steps are the necessary prerequisites to facilitate reuse, but for reuse to be achieved it is imperative that the linguistics and semantic web community collaborate and share their expertise.

References

Bizer C, Heath T, Berners-Lee T (2009a) Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) 5(3):1–22

Bizer C, Lehman J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009b) DBpedia - A crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3):154–165

Chiarcos C (this vol.) Interoperability of corpora and annotations. P. 161-179

Chiarcos C, Hellmann S, Nordhoff S (this vol.) The Open Linguistics Working Group of the Open Knowledge Foundation. P. 153-160

Consortium LD (2005) ACE (Automatic Content Extraction) English Annotation Guidelines for Entities version 5.6.1

Eckart K, Riester A, Schweitzer K (this vol.) A discourse information radio news database for linguistic analysis. P. 65-75

Euzenat J, Meilicke C, Stuckenschmidt H, Shvaiko P, Trojahn C (2011) Ontology alignment evaluation initiative: Six years of experience. Journal on Data Semantics 15:158–192

Fellbaum C (ed) (1998) WordNet: An Electronic Lexical Database. The MIT Press

Halpin H, Hayes PJ, McCusker JP, McGuinness DL, Thompson HS (2010) When owl:sameAs isn't the same: An analysis of identity in linked data. In: The 9th International Semantic Web Conference (ISWC 2010), Shanghai, China, pp 305–320

Hellmann S, Stadler C, Lehmann J (this vol.) The German DBpedia: A sense repository for linking entities. P. 181-189

Mika P, Ciaramita M, Zaragoza H, Atserias J (2008) Learning to tag and tagging to learn: A case study on Wikipedia. IEEE Intelligent Systems 23(5):26–33

Petrov S, Das D, McDonald R (2011) A universal part-of-speech tagset. arXiv:1104.2086v1

Pustejovsky J, Lee K, Bunt H, Romary L (2010) ISO-TimeML: An international standard for semantic annotation. In: Proceedings of LREC 2010, pp 394–397

Recasens M, Hovy E, Martí MA (2011) Identify, non-identity, and near-identity: Addressing the complexity of coreference. Lingua pp 1138–1152

Teufel S (1997) A support tool for tagset mapping. arXiv:cmp-lg/9506005v2

Tjong Kim Sang EF, Meulder FD (2003) Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada, pp 142–147

Windhouwer M, Wright SE (this vol.) Linking to linguistic data categories in ISOcat. P. 99-107