FrameNet and Linked Data

Nancy Ide
Department of Computer Science, Vassar College
Poughkeepsie, New York USA
ide@cs.vassar.edu

Abstract

FrameNet is the ideal resource for representation as linked data, and several renderings of the resource in RDF/OWL have been created. FrameNet has also been and continues to be linked to other major resources, including WordNet, BabelNet, and MASC, in the Linguistic Linked Open Data cloud. Although so far the supporting technologies have not enabled easy and widespread access to the envisioned massive network of language resources, a conflation of recent efforts suggests this may be a reality in the not-too-distant future.

FrameNet (Fillmore et al., 2002; Ruppenhofer et al., 2006) is the ideal resource for representation in the Semantic Web (SW) as what is now widely known as "linked data". The Semantic Web consists of objects whose properties are represented by named links to other objects that constitute their values and supports representing and reasoning over ontologies defined the the SW framework. FrameNet is also a complex semantic network linking lexical units to semantic frames, and semantic frames to one another in a shallow hierarchy, over which inheritance and sub-frame relations are defined. In sentences annotated for FrameNet frame elements, the role is a property of a frame object that is linked to the entity (object) that fills it; FrameNet also includes a hierarchy of semantic types that constrain the possible fillers for a given role. FrameNet thus defines a dense network of objects and properties supported by ontological relations–exactly what the Semantic Web is intended to be.[1]

The suitability of FrameNet for representation in the Semantic Web was recognized fairly early on in the development of the family of Semantic Web formats, which include the Resource Definition Framework (RDF) and the Web Ontology Language (OWL), which first became available as W3C standards in the late 90s and early 2000s. In one of the earliest projects to adapt linguistic resources to the Semantic Web, FrameNet was rendered in RDF and DAML+OIL (the precursor to OWL) in 2003, soon after these formats first became standardized, for the stated goal of providing "a potential resource to aid in the automatic identification and disambiguation of word meanings on the semantic web" (Narayanan et al., 2003a). Later, the DAML+OIL portion was converted to OWL (Scheffczyk et al., 2008; Scheffczyk et al., 2010). Other conversions include (Coppola et al., 2009) and (Narayanan et al., 2003b); most recently, FrameNet was ported to RDF/OWL for inclusion in the Linked Open Data (LOD) cloud[2] (Nuzzolese et al., 2011). The possibility of linking WordNet and FrameNet in the Semantic Web has also spawned efforts such as (Bryl et al., 2012) that build on numerous efforts over the past several years to align and/or extend these two resources (Burchardt et al., 2005; Ide, 2006; De Cao et al., 2008; de Melo et al., 2012; Bryl et al., 2012). Others have analyzed FrameNet in order to formalize its semantics so as to be appropriate for use with Description Logic (DL) reasoners compatible with OWL-DL (Ovchinnikova et al., 2010).

Given all of the activity surrounding FrameNet as a resource for the Semantic Web, one would expect to see multiple examples of the use of Semantic Web implementations of FrameNet for NLP development and research. However, these examples do not exist, for two reasons. The first is a reality of the Semantic Web: simply put, the Semantic Web has not yet come to fruition, despite its having been around as a concept for well over a decade, and despite the development of a suite of W3C standard technologies to support it.

[1] For a fuller description of the structure of FrameNet data, see (Scheffczyk et al., 2008).

[2] http://linkeddata.org

Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014), pages 18–21,
Baltimore, Maryland USA, June 27, 2014.

One of the most important of these technologies is SPARQL (Prud'hommeaux and Seaborne, 2008), a query language for data in RDF format, which is the crucial tool for exploiting the inter-linkages among linguistic resources to support NLP. Unfortunately, SPARQL is new enough that it is not yet widely deployed and has not had the benefit of decades of optimization to improve its performance, which so far often suffers from sluggishness. The good news is that new research and implementations are rapidly contributing to the improvement of SPARQL and other Semantic Web technologies, and as a result, we are seeing signs that the requisite base infrastructure may be (or may soon be) in place to support accelerated growth and deployment.

At the same time, over the past four or five years several efforts in Semantic Web development–in particular, the deployment of knowledge bases, lexicons, ontologies, and similar resources as linked data–have made notable progress, including the LOD cloud and, of special interest for the NLP community, its companion Linguistic Linked Open Data (LLOD) cloud (Chiarcos et al., 2012a). Efforts to link, especially, lexical-semantic databases like FrameNet, WordNet, and BabelNet (Navigli and Ponzetto, 2010) are underway, although full, operational linkage may not be immediate. At the same time, however, there is virtually no language *data* in the form of corpora in the LLOD, and none that contains annotations that can be linked to lexicons and knowledge bases.

This suggests a second reason why FrameNet as linked data has not yet been used in NLP research: a more useful configuration for a FrameNet-based resource in the Semantic Web would include linkage from frame governors and frame elements to (many) examples in corpora, rather than a simple rendering of linkages among lexical units, frames, and frame elements. Coupled with linkage to WordNet and multilingual semantic resources such as BabelNet (which has also been recently ported to RDF–see (Navigli, 2012)), a Semantic Web resource of this type and magnitude would enable SPARQL queries that collect information across several linguistic phenomena and levels, for example, "find all tokens in English and Russian that refer to *land* as a political unit (synonyms from the WordNet synset land%1:15:02::) in the VICTIM role of the ATTACK frame". This is a trivial example; the full range of possibilities is left to the reader's imagination, and awaits SPARQL's transition to full adulthood.

FrameNet has always hand-annotated sample sentences as input to frame construction, due to the insistence by FrameNet's founder on grounding the theory in real language data. FrameNet's early annotation efforts used examples from the British National Corpus (BNC); however, as time went on, FrameNet and similar annotation projects[3] found that usage examples extracted from the BNC were often unusable or misrepresentative for developing templates to describe semantic arguments and the like, due to significant syntactic differences between British and American English. This motivated a proposal for an American National Corpus (ANC)[4] (Fillmore et al., 1998), comparable to the BNC but including genres nonexistent at the time of BNC development (blogs, email, chat rooms, tweets, etc.) as well as annotations beyond part-of-speech, to serve as basis for the development of lexical-semantic resources and NLP research in general.[5]

In 2006, the ANC, FrameNet, and WordNet projects received a substantial grant from the U.S. National Science Foundation[6] to produce a half-million word Manually Annotated Sub-Corpus (MASC)[7] (Ide et al., 2010), consisting of data drawn from the ANC and annotated for multiple types of linguistic phenomena. The project included a component to annotate portions of the corpus for WordNet senses and FrameNet frame elements, in order to provide input to an effort to harmonize these two resources (Baker and Fellbaum, 2009). The first full version of the corpus, released in 2012, included over 16 different annotation types and was coupled with a separate sentence corpus (Passonneau et al., 2012) that includes WordNet 3.1 sense-tags for approximately 1000 occurrences of each of 114 words chosen by the WordNet and FrameNet teams (ca. 114,000 annotated occurrences). Of these, 100 occurrences of each word (over 1000 sentences) are also anno-

[3]E.g., Comlex (http://nlp.cs.nyu.edu/comlex/) and NomLex (http://nlp.cs.nyu.edu/nomlex/)

[4]http://www.anc.org/

[5]The ANC never received the substantial funding and text contributions enjoyed by the BNC, and as a result has so far released only 22 million words of data, including a 15 million word subset that is unrestricted for any use called the Open ANC" (OANC), available at http://www.anc.org/data/oanc/.

[6]NSF CRI 0708952

[7]http://www.anc.org/data/masc/

tated for FrameNet frame elements. These annotations were subsequently used in a major WordNet-FrameNet alignment effort (de Melo et al., 2012).

MASC provides a missing link in the Semantic Web scenario for FrameNet and related resources. The corpus contains not only FrameNet and WordNet annotations, but also annotations over parts or all the corpus at several other linguistic layers including morphosyntax, syntax (shallow parse, Penn Treebank annotation), semantics (named entities, opinion, PropBank), and discourse (coreference, clause boundaries and nucleus/satellite relations). All of MASC is currently being incorporated into the LLOD cloud, and its FrameNet and WordNet annotations will be linked to the linked data versions of those resources.[8] The resulting resource, connecting multiple major semantic resources and a broad-genre corpus, will be unparalleled as a foundation for NLP research and development.

When the annotations for other phenomena in MASC are added into the mix, the potential to study and process language data *across* multiple linguistic levels becomes even greater. It is increasingly recognized that to perform human-like language understanding, NLP systems will ultimately have to dynamically integrate information from all linguistic levels as they process input, but despite this recognition most work in the field continues to focus on isolated phenomena or utilizes only selected phenomena from a few linguistic levels. Some corpora with multiple annotation layers, including MASC and a (very few) others such as OntoNotes (Pradhan et al., 2007), have recently been created, but due to the high costs of their development they are limited in size and do not include annotations across the gamut of linguistic phenomena. Similarly, standardized formats for annotated data (e.g., (ISO, 2012)), lexical-semantic resources (ISO, 2008), and reference categories for linguistic annotations (Marc Kemps-Snijders and Wright, 2008) have been developed to enable merging of annotations of different types and formats, but there still remains considerable disparity among and/or lack of support for processing merged resources.

[8] See (Chiarcos et al., 2012b) for a discussion of the process and benefits. BabelNet annotations of MASC, which are in turn linked to wordnets in multiple languages, have also been recently contributed (Moro et al., 2014), thus opening up the possibility for linkage from MASC to that resource as well–and, by extension, linkage between BabelNet and MASC's existing FrameNet and WordNet annotations.

Is the Semantic Web the answer? Will it be the vehicle for a paradigm shift in NLP research and development? Likely, it or something it evolves into will ultimately provide the required common representation and processing framework which, coupled with potentially enormous advances in computer and network speed, data capacity, neurotechnology, network-on-chip technologies, and the like, will fundamentally change our approach to NLP in the decades to come. In the meantime, it remains to be seen how quickly Semantic Web technology will progress, and how soon most or all language resources will reside in places like the LLOD cloud, so that they can begin to be fully and readily exploited. But whether the Semantic Web as we know it now is the ultimate solution or simply a developmental step, the direction seems clear; and fittingly, FrameNet is one of the first resources on board.

References

Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 125–129.

Volha Bryl, Sara Tonelli, Claudio Giuliano, and Luciano Serafini. 2012. A novel Framenet-based resource for the semantic web. In *SAC*, pages 360–365.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet detour to FrameNet. In *Proceedings of the GLDV 2005 workshop GermaNet II*.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012a. Linking Linguistic Resources: Examples from the Open Linguistics Working Group. In *Linked Data in Linguistics*, pages 201–216. Springer.

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2012b. Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer.

Bonaventura Coppola, Aldo Gangemi, Alfio Massimiliano Gliozzo, Davide Picca, and Valentina Presutti. 2009. Frame Detection over the Semantic Web. In *Proceedings of the 6th European Semantic Web Conference*.

Diego De Cao, Danilo Croce, Marco Pennacchiotti, and Roberto Basili. 2008. Combining Word Sense and Usage for Modeling Frame Semantics. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 85–101.

Gerard de Melo, Collin F. Baker, Nancy Ide, Rebecca Passonneau, and Christiane Fellbaum. 2012. Empirical Comparisons of MASC Word Sense Annotations. In *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Charles J. Fillmore, Nancy Ide, Daniel Jurafsky, and Catherine Macleod. 1998. An American National Corpus: A Proposal. In *Proceedings of the First Annual Conference on Language Resources and Evaluation*, pages 965–969, Granada, Spain.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. The FrameNet Database and Software Tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, volume IV.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus: A Community Resource for and by the People. In *Proceedings of ACL 2010*, pages 68–73.

Nancy Ide. 2006. Making Senses: Bootstrapping Sense-Tagged Lists of Semantically-Related Words. In *Computational Linguistics and Intelligent Text*, pages 13–27.

2008. Language Resource Management – Lexical Markup Framework. International Standard ISO 24613.

2012. Language Resource Management – Linguistic Annotation Framework. International Standard ISO 24612.

Peter Wittenburg Marc Kemps-Snijders, Menzo Windhouwer and Sue Ellen Wright. 2008. ISOCat: Corralling Data Categories in the Wild. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Andrea Moro, Roberto Navigli, Francesco Maria Tucci, and Rebecca J. Passonneau. 2014. Annotating the MASC corpus with babelnet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC' 2014)*.

Srini Narayanan, Collin F. Baker, Charles J. Fillmore, and Miriam R.L. Petruck. 2003a. FrameNet Meets the Semantic Web: Lexical Semantics for the Web. In *The Semantic Web - ISWC 2003*, pages 771–787. Springer.

Srinivas Narayanan, Miriam R.L. Petruck, Collin F. Baker, and Charles J. Fillmore. 2003b. Putting FrameNet Data into the ISO Linguistic Annotation Framework. In *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, page 22–29.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225.

Roberto Navigli. 2012. BabelNet goes to the (Multilingual) Semantic Web. In *ISWC 2012 Workshop on Multilingual Semantic Web*.

Andrea Giovanni Nuzzolese, Aldo Gangemi, and Valentina Presutti. 2011. Gathering lexical linked data and knowledge patterns from FrameNet. In *K-CAP*, pages 41–48.

Ekaterina Ovchinnikova, Laure Vieu, Alessandro Oltramari, Stefano Borgo, and Theodore Alexandrov. 2010. Data-Driven and Ontological Analysis of FrameNet for Natural Language Reasoning. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Rebecca J. Passonneau, Collin F. Baker, Christiane Fellbaum, and Nancy Ide. 2012. The MASC Word Sense Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Sameer S. Pradhan, Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 517–526.

Eric Prud'hommeaux and Andy Seaborne. 2008. SPARQL Query Language for RDF.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute.

Jan Scheffczyk, Collin F. Baker, and Srini Narayanan. 2008. Ontology-Based reasoning about lexical resources. In *Ontologies and Lexical Resources for Natural Language Processing*. Cambridge University Press.

Jan Scheffczyk, Collin Baker, and Srrini Narayanan, 2010. *Reasoning over Natural Language Text by Means of FrameNet and Ontologies*, pages 53–71. Cambridge University Press.