SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Abstracts of the Dagstuhl Seminar on the "Multilingual Semantic Web"

September 2nd - 7th 2012

Schloss Dagstuhl, Wadern

Organizers: Paul Buitelaar, Philipp Cimiano, Ed Hovy and Key-Sun Choi

Dagstuhl seminar on the Multilingual Semantic Web

Short position statement

Guadalupe Aguado de Cea and Elena Montiel-Ponsoda (Universidad Politécnica de Madrid)

1) Most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web (SW):

Many attempts have been made to provide multilinguality to the Semantic Web, by means of annotation properties in Natural Language (NL), such as RDFs or SKOS labels and other lexicon-ontology models, such as *lemon*, but there are still many issues to be solved if we want to have a truly accessible Multilingual Semantic Web (MSW). **Reusability of monolingual resources** (ontologies, lexicons, etc.), **accessibility of multilingual resources hindered by many formats**, **reliability of ontological sources**, **disambiguation problems** and **multilingual presentation to the end user of all this information in NL** can be mentioned as some of the most relevant problems. Unless this NL presentation is achieved, MSW will be restricted to the limits of IT experts, but even so, with great dissatisfaction and disenchantment.

2) Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?

Considering Linked Data as a step forward from the original Semantic Web, to provide the possibility of accessing all the information gathered in all the ontological resources should become one significant objective, if we want every user to “perform searches in their own language”, as mentioned in the motivation of Dagstuhl Seminar.

Globalization of work has opened the scope of possible domains and sectors interested in Linked data and a true MSW. From governmental, political, administrative, economic issues to medicine, chemistry, pharmaceutical, car makers and other industries alike, all would hop on the bandwagon of MSW if it provides them the suitable information needed for their businesses.

As long as we cannot retrieve the answer to a question in NL, even if we have the possible information in DBpedia, and other ontological and knowledge resources, it will be difficult to beat Google, and extract the most of LD and the SW, no matter how many “semantic” resources we have.

3) Which figures are suited to quantify the magnitude or severity of the problem?

It is difficult for us to quantify the problem in figures, but it is clear that if this issue remains unsolved, we can miss the boat. In the last few years the mobile industry has made advances at a greater speed, maybe because there were more chances to make money.

4) Why do current solutions fail short?

At the moment, we have complex models to be implemented by SW illiterate, many technological issues unsolved, lack of agreement with respect to the ontological-lexical linguistic knowledge to be provided to end-users when using the SW to improve their resources

5) What insights do we need in order to reach a principled solution? What could a principled solution look like?

Focusing on certain aspects that can be agreed upon by many key sectors (researchers, developers, industry, end-users) some relevant problems could be approached aiming at delimiting the wishes, needs and resources available. A principled solution should be based on simplicity, usefulness, wide coverage, and reusability

6) How can standardization (e.g. by the W3C) contribute?

It can contribute because participation is open to many sectors involved. If all sectors cooperate, dissemination and promotion can be achieved more easily. Getting other standardization committees involved (ISO TC 37) can also widen the scope and can contribute too to dissemination. But it is important to get industry professionals involved to make them aware of the possibilities they have to make the most of their products.

Accessibility to a Pervasive Web for the challenged people

Dimitra Anastasiou

1) What are in your view the most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web?

One need is to make people believe about its importance. Although some projects and workshops (including the Dagstuhl Workshop) bring forward this topic, still there is need for more interesting projects and initiatives in the community. As Semantic Web technologies are used by many domains and multilingualism is also an aspect taken into account by many stakeholders, many people regard the Multilingual Semantic Web (MSW) as a vague concept, so some clear description, specifications or even a standard would make the MSW more prominent.

I am more interested at the moment in the accessibility to the Web and the MSW by the seniors and people with disabilities. Moreover, and in relation to the Web for challenged people, I am interested in the pervasive Web in Ambient Assisted Living (AAL), which goes beyond the Web present on a PC monitor, and is present in the invisible technology in smart homes.

2) Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?

The ageing phenomenon is reality today, as according to the World Population Ageing report, the world average of the 65+ age group was 7.6% in 2010 and will be 8.2% in 2015.

The European Commission suggests demographic and epidemiological research on ageing and disability, predicting the size of the future ageing population, and acquiring information as inputs to planning. Mostly industries (and some academic groups) are concerned with AAL, but the community researching on the Web technology used particularly there is very small. Moreover, multilingualism plays a secondary role, though it is so important, as seniors today are often not foreign language speakers and have to communicate with technology (Web or not). Whereas health informatics, HCI, HRI, sensoring and recognition play an important role, the Semantic Web and multilingual support are not taken into serious consideration.

2) Which figures are suited to quantify the magnitude or severity of the problem?

The Working Draft of “Web Accessibility for Older Users: A Literature Review“ gives very interesting insights about Web design and development and its aspects affecting the elderly.

3) Why do current solutions fail short?

Because the limitations of those challenged people can vary significantly, cannot be really categorized in specific groups, so high customization of software and high learning effort is needed which results in information overload. The technology is not affordable yet, but rather too expensive. Moreover, it is also very complex, so easier-to-use and user-friendly methods should be developed.

4) What insights do we need in order to reach a principled solution? What could a principled solution look like?

More initiatives including common projects, community groups workshops in the fields of AAL, multimodality, semantic Web, language technology.
A principled solution should look like elderly persons being able to speak in their mother tongue to to turn on and off their coffee machine, switch on and off lights. When they speak in their mother tongue, they do not feel digitally intimidated, but are more natural, trustful, and user-friendly. Ontologies could help dialogue systems triggering predictable actions in AAL smart homes, i.e. turning off the oven when not used or reminding a person to make a phone call.

5) How can standardization (e.g. by the W3C) contribute?

Cooperation with the W3C Web Accessibility Initiative would be very useful. It has released Web Content Accessibility Guidelines, User Agent Accessibility Guidelines, and Authoring Tool Accessibility Guidelines.

[1] W3C Web Accessibility Initiative (WAI): http://www.w3.org/WAI/

[2] Web Content Accessibility Guidelines (WCAG) Overview: http://www.w3.org/WAI/intro/wcag.php

[3] Web Accessibility for Older Users: A Literature Review: http://www.w3.org/TR/wai-age-literature/

Position Paper:
Dagstuhl Seminar on Multilingual Semantic Web:

Multilingual Computation with Resource and Process Reusage

Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT Bombay, India
pb@cse.iitb.ac.in

1 Introduction

Mutilingual computation is the order of the day and is needed critically for the realization of the semantic web dream. Now, it stands to reason that *work done* for a language should come to help for computation in another language. For example, if through the investment of resources we have been able to detect named entities in one language, we should be able to detect them in another language too, through much smaller level of investment like transliteration. The idea of **projection** from one language to another is a powerful and potent one and merits deep investigation. In the seminar I would like to expound on the *projection for multilingual NLP*.

2 Challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web

Resource constraint is the most important challenge facing multilingual access to the semantic web. Over the years through conscious decisions, English has built foundational tools and resources for language processing. Examples of these are Penn Treebank[1], Propbank, Rule based and Statistical Parsers[2], Wordnet[3], Corpora of various kinds of annotation and so on and so forth. No language comes anywhere close to English in terms of lexical resources and tools.

3 Why does the problem matter in practice?

It is impossible to do NLP without adequate lexical resources and foundational tools. For example, nobody thinks of building a parser today for a language, without first creating Treebank for the language- constituency or dependency- and then training a probabilistic parser on the treebank. However, creating treebanks requires years of effort.

Everything in language technology sector needs lexical resources. Information Extraction, Machine Translation and Cross Lingual Search are some of the examples. E-Governance- a domain dealing with the automatization of administrative processes of the Government in a large, multilingual country like India- is a large consumer of language technology.

4 Which figures are suited to quantify the magnitude or severity of the problem?

Lexical resources are typically quantified by the amount of annotated data and foundational tools by their precision and recall figures. For example, the famous SemCor[4] corpus for sense

[1] http://www.cis.upenn.edu/~treebank/
[2] http://nlp.stanford.edu/software/lex-parser.shtml
[3] http://wordnet.princeton.edu/
[4] http://www.gabormelli.com/RKB/SemCor_Corpus

annotated data has about 100,000 wordnet id marked words. On the tools side, CLAWS POS tagger for English has over 97% accuracy.

5 Why do current solutions fail short?

It takes years to build high quality lexical resources. Both linguistic expertise and computational dexterity are called for. It is not easy to find people with both linguistic and computational acumen. Large monetary investment to is called for.

6 Principled Solution

Projection is the way to go. Reuse of resources and processes is a must. Over the years in our work on word sense disambiguation involving Indian languages, we have studied how sense distributions can be projected from one language to another for effective WSD (see bibliography below). The idea of projection has been applied in POS tagging (best paper award ACL 2011[5]). We have also used it to learn named entities in one language from the NE tagged corpora of another language.

7 How can standardization (e.g. by the W3C) contribute?

For projection to work at all, resources and tools need to be standardized for input-output, storage, API and so on. For example, wordnet building activity across the world follows the standard set by the Princeton WordNet.

Bibliography

1. Mitesh Khapra, Salil Joshi and Pushpak Bhattacharyya, *It takes two to Tango: A Bilingual Unsupervised Approach for Estimating Sense Distributions using Expectation Maximization*, 5th International Joint Conference on Natural Language Processing (**IJCNLP 2011**), Chiang Mai, Thailand, November 2011.

2. Mitesh Khapra, Salil Joshi, Arindam Chatterjee and Pushpak Bhattacharyya, *Together We Can: Bilingual Bootstrapping for WSD*, Annual Meeting of the Association of Computational Linguistics (**ACL 2011**), Oregon, USA, June 2011.

3. Mitesh Khapra, Saurabh Sohoney, Anup Kulkarni and Pushpak Bhattacharyya, *Value for Money: Balancing Annotation Effort, Lexicon Building and Accuracy for Multilingual WSD*, Computational Linguistics Conference (**COLING 2010**), Beijing, China, August 2010.

4. Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya, *Projecting Parameters for Multilingual Word Sense Disambiguation, Empirical Methods in Natural Language Processing (**EMNLP09**), Singapore, August, 2009.*

[5] Dipanjan Das and Slav Petrov, *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections (**ACL11**), Singapore, August, 2009*

Notes on Multilingual Semantic Web and the challenges of Open Language Data –

Open Language Resources & Meta-Resources & Open Research Results

Nicoletta Calzolari

Language Technology (LT) is a data-intensive field and major breakthroughs have stemmed from a better use of more and more Language Resources (LRs). LRs and Open/Shared Language Data is therefore a great topic! New approaches are needed, both for Data and Meta-Data (LRs and Meta-LRs). My topics are linked to the layer of LRs and language services that serve LT, and especially open information on LRs and on research results. How can Linked Data contribute?

1. The Language Resource dimensions

In the FLaReNet Final Blueprint, the actions recommended for a strategy for the future of the LR field are organised around nine dimensions: a) *Infrastructure*, b) *Documentation*, c) *Development*, d) *Interoperability*, e) *Coverage, Quality and Adequacy*, f) *Availability, Sharing and Distribution*, g) *Sustainability*, h) *Recognition*, i) *International Cooperation*. Taken together, as a coherent system, these directions contribute to a *sustainable LR ecosystem*.

Multilingual Semantic Web has strong relations with many of these dimensions, esp. a), b), d), f), g).

2. Language Resources and the Collaborative framework

The traditional LR production process is too costly. A new paradigm is pushing towards open, distributed language infrastructures based on sharing LRs, services and tools. It is urgent to create a framework enabling effective cooperation of many groups on common tasks, adopting the paradigm of accumulation of knowledge so successful in more mature disciplines, such as biology, astronomy, physics. This requires the design of a new generation of multilingual LRs, based on open content interoperability standards.

Multilingual Semantic Web may help in determining the shape of the LRs of the future, consistent with the vision of an open distributed space of sharable knowledge available on the web for processing. It may be crucial to the success of such an infrastructure, critically based on interoperability, aimed at improving sharing of LRs and accessibility to multilingual content. This will serve better the needs of language applications, enabling building on each other achievements, integrating results, and having them accessible to various systems, thus coping with the need of more and more ‘knowledge intensive’ large-size LRs for effective multilingual content processing. This is the only way to make a great leap forward.

3. Open Documentation on LRs

Accurate and reliable documentation of LRs is an undisputable need: documentation is the gateway to discovery of LRs, a necessary step towards promoting the data economy. LRs that are not documented virtually do not exist: initiatives able to collect and harmonise metadata about resources represent a valuable opportunity for the NLP community.

LRE Map

The LRE Map is a collaborative bottom-up means of collecting metadata on LRs from authors. It is an instrument for enhancing availability of information about LRs, either new or already existing ones, and a way to show the current LR landscape and its trends. As a measuring tool for monitoring various dimensions of LRs across places and times, it helps highlighting evolutionary trends in LR use and related development by cataloguing not only LRs in a narrow sense (i.e.

language data), but also tools, standards, and annotation guidelines. The Map contributes to the promotion of a movement towards an accurate and massive documentation of LRs.

4. Open Language Resource Repositories

The rationale behind the need of Open LR Repositories is that accumulation of massive amounts of (high-quality) multi-dimensional data about many languages is the key to foster advancement in our knowledge about language and its mechanisms. We must be coherent and take concrete actions leading to the coordinated gathering − in a shared effort − of as many (processed/annotated) language data as we are able to produce. This initiative compares to the astronomers/ astrophysics' accumulation of huge amounts of observation data for a better understanding of the universe.

Language Library

The Language Library is an experiment – started around parallel/comparable texts processed by authors at LREC 2012 – of a facility for gathering and making available the linguistic knowledge the field is able to produce, putting in place new ways of collaboration within the community. It is collaboratively built by the community providing/enriching LRs by annotating/processing language data and freely using them. The multi-layer and multi-language annotation on the same parallel/comparable texts should foster comparability and equality among languages.

The Language Library is conceived as a theory-neutral space, which allows for several annotation philosophies to coexist, but we must exploit the sharing trend for initiating a movement towards creating synergies and harmonisation among annotation efforts that are now dispersed. In a mature stage the Library could focus on enhancing interoperability, encouraging the use of common standards and schemes of annotation. Interoperability should not be seen as a superimposition of standards but rather as the promotion of a series of best practices that might help other contributors to better access and easily reuse the annotation layers provided. The Language Library could be seen as the beginning of a big Genome project for languages, where the community collectively deposits/creates increasingly rich and multi-layered LRs, enabling a deeper understanding of the complex relations between different annotation layers/language phenomena.

5. Open Repositories of Research Results

Disclosing data/tools related to published papers is another "simpler" addition to the Language Library, contributing to the promotion of open repositories of LR research results. Moreover LRs must be not only searchable/shareable, but also "citable" (linked to issue h) "*recognition*").

6. Open Language Data (OpenLanD)

Open Language Data – the set of 2. to 5. above – aims at offering the community a series of facilities for easy and broad access to information about LRs in an authoritative and trustable way. By investing in data reusability, OpenLanD can store the information as a collection of coherent datasets compliant to the Linked Data philosophy. The idea is that by linking these data among themselves and by projecting them onto the wider background of Linked Data, new and undiscovered relations can emerge. OpenLanD must be endowed with functionalities for data analytics and smart visualisation. OpenLanD differs from existing catalogues for the breadth and reliability of information due to a community-based approach. The information made available covers usages, applications of LRs, their availability, as well as related papers, individuals, organisations involved in creation or use, standards and best practices followed or implemented. OpenLanD avoids the problem of rapid obsolescence of other catalogues by adopting a bottom-up approach to meta-data population.

On multilingual web sites

M.T. Carrasco Benitez
European Commission, Luxembourg, July 2012, version 1.0

1. Abstract

Multilingual Web Sites (MWS) refers to web sites that contain multilingual parallel texts; i.e., texts that are translations of each other. For example, most of the European Institutions sites are MWS, such as Europa [EU]. The main point of views are:

- **Users** should expect the same multilingual behaviour when using different browsers and/or visiting different web sites.

- **Webmasters** should be capable of creating quickly high quality, low cost MWS.

This is a position paper for the Dagstuhl Seminar on the Multilingual Semantic Web. Personal notes on this event are at:

http://dragoman.org/dagstuhl

2. Relevance

MWS are of great practical relevance as there are very important portals with many hits; also they are very complex and costly to create and maintain: Europa is in 23 languages and contains over 8 million pages. Facilitating and enjoying this common experience entails **standardisation**: current multilingual web sites are *applications* incompatible with each other. There is a Multilingual Web Sites Community Group at the W3C [CG].

3. Point of views

3.1. User

From a users point of view, the most common usage is **monolingual**, though a site might be multilingual; i.e., users are usually be interested in just one language of the several available at the server. The language selection is just a barrier to get the appropriate linguistic version. One has also to consider that some users might be really interested in several linguistic versions. It is vital to agree on common behaviours for users: browser-side (*language button*) and server-side (*language page*).

3.1. Webmaster

Webmaster refers to all the aspect of the construction of MWS: author, translator, etc. The objective is the creation of high quality low cost MWS. Many existing applications have some multilingual facilities and (stating the obvious) one should harvest the best techniques around.

Servers should expect the same application programming interface (API). The first API could be just a multilingual data structure. The absence of this data structure means that each application has to craft this facility; having the same data structure means that servers (or other programs) would know how to process this data structure directly. It is a case of production of multilingual parallel texts: the cycle *Authorship, Translation and Publication chain* (ATP-chain) [MPT].

4. Wider context

- **Language vs. non-language aspects**: differentiate between aspects that are language and non-language specific. For example, the API between CMS and web server is non-language specific and it should be addressed in a different forum.
- **Language as a dimension**: as in TCN, one should consider language a *dimension* and extend the concept to other areas such as linked data. Consider also *feature negotiations* as in TCN.
- **Linguistic versions**: the *speed* (available now or later) and translation *technique* (human or machine translation) should be considered in the same model.
- **Unification**: multilingual web is an exercise in unifying different traditions looking at the same object from different angles and different requirements. For example, the requirements for processing a few web pages are quite different from processing a multilingual corpus of several terabytes of data.

4. Multidiscipline map

- Web technology proper
  - Content management systems (CMS), related to authoring and publishing
  - Multilingual web site (MWS)
  - Linked data, a form of multilingual corpora and translation memories
  - Previous versions in time, a form of archiving [MEMENTO]
- Traditional natural language processing (NLP)
  - Multilingual corpora, a form of linked data [MUSET]
    - Source documents and tabular transformations, the same data in different presentations
  - Machine translation, for end users and prepossessing translators
- Translation
  - Computer-aided translation (CAT)
    - Preprocessing, from corpora, translation memories or machine translation
  - Computer-aided authoring, as a help to have better source text for translation
  - Localisation
  - Translation memories (TM) [TMX], related to corpora and linked data
- Industrial production of multilingual parallel publications
  - Integration of the Authorship, Translation and Publishing chain (ATP-chain)
  - Generation of multilingual publications
  - Official Journal of the European Union [OJ]

5. Disclaimer

This document represents only the views of the author and it does not necessarily represent the opinion of the European Commission.

6. References

[CG] Multilingual Web Sites Community Group; http://www.w3.org/community/mws
[EU] Europa; http://europa.eu
[MEMENTO] Memento - Adding Time to the Web; http://mementoweb.org
[MPT]Open architecture for multilingual parallel texts; http://arxiv.org/pdf/0808.3889
[MUSET] Multilingual Dataset Format; http://dragoman.org/muset
[PN] Personal notes for this event; http://dragoman.org/dagstuhl
[OJ] Official Journal of the European Union; http://publications.europa.eu/official/index_en.htm
[TCN] Transparent Content Negotiation in HTTP; http://tools.ietf.org/rfc/rfc2295.txt
[TMX] TMX 1.4b Specification; http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

Dagstuhl Seminar on the Multilingual Semantic Web, Position Paper

Christian Chiarcos[1]

The premise of the Dagstuhl seminar is the question which problems we need to overcome in order to enhance multilingual access to the Semantic Web, and how these are to be addressed.

Ultimately, the Semantic Web in its present stage suffers from a predominance of resources originating in the Western hemisphere, with English as their primary language. Eventually, this could be overcome by providing translated and localized versions of resources in other languages, and thereby creating a critical mass of foreign language resources that is sufficient to convince potential non-English speaking users to (a) employ these resources, and (b) to develop their own extensions or novel resources that are linked to these. On a large scale, this can be done automatically only, comparable to, say, the conversion of the English Wikipedia into Thai.[2] Unlike the translation of plain text, however, this translation requires awareness to the conceptual structure of a resource, and is thus not directly comparable to text-oriented Machine Translation. A related problem is that the post-editing of translation results in a massive crowdsourcing approach (as conducted for the Thai Wikipedia) may be problematic, because most laymen will not have the required level of technical understanding.

Therefore, the task of resource translation (and localization) of Semantic Web resources requires a higher level of automated processing than comparable amounts of plain text. This is an active research topic, but pursued by a relatively small community. One possible issue here is that the NLP and Semantic Web communities are relatively isolated from each other,[3] so that synergies between them are limited. A consequence is that many potentially interested NLP people are relatively unaware of developments in the Semantic Web community, and, moreover, that they do not consider Semantic Web formalisms to be relevant to their research. This is not only a problem for the progress of the Multilingual Semantic Web, but also for other potential fields of overlap. In the appendix, I sketch two of them.

In my view, both the NLP community and the Semantic Web community could benefit from small- to mid-scale events co-located with conferences of the other community (or joint seminars, as this workshop), and that this may help to identify fields of mutual interest, including, among other topics, the translation of Semantic Web resources. In at least two other fields, such convergence processes may already be underway, as sketched below.

[1] Christian Chiarcos, chiarcos@daad-alumni.de, Information Sciences Institute, University of Southern California

[2] http://www.asiaonline.net/portal.aspx#ThaiLaunch

[3] For example, the LREC (http://www.lrec-conf.org) lists 11 publications for the topic „Semantic Web“ for 2012,11 for 2010, 16 for 2008. Similarly, the Google counts for ACL (http://aclweb.org/anthology) contributions containing the word „ontology“ are consistently low: 2008: 5, 2009: 8, 2010: 15, 2011: 3, 2012: 7. Both conferences have between 500 and 1000 participants, so, in terms of paper-participant ratio, this line of research is underrepresented.

Questionnaire

1) Challenges/ problems and needs with respect to the multilingual access to the Semantic Web

For languages that are under-represented in the Semantic Web, the initial bias to create resources in their own language and in accordance with their own culture is substantially higher than for English, where synergy effects with existing resources can be exploited in the development of novel resources. To provide these languages with a basic repository of SW resources, massive automated translation is required. This task is, however, closer to the traditional realm of NLP than to that of the SW. The SW-subcommunity working towards this direction is thus relatively small, and may benefit from closer ties to the NLP community. (Which may be of mutual interest to both sides, also beyond the problem of Semantic Web multilingualism, see appendix.)

2) Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?

The situation is comparable to the development of NLP tools for less-resourced languages. Without a basic set of language- and culture-specific resources (say, a WordNet and a DBpedia/Wikipedia with sufficient coverage), there will be little interest to develop and to invest in Semantic Web applications. A plain translation is an important first step, but for semantic resources, there may be important culture-specific differences that need to be taken into consideration. These efforts can be crowd-sourced to a certain extent, but only if a certain level of knowledge is already available in order to convince contributors that this is an effort that pays off.

3) Which figures are suited to quantify the magnitude or severity of the problem?

As for the primary problem to attract potentially interested NLP people, this can be illustrated by the small number of Semantic Web contributions to NLP conferences (and vice versa), see footnote 3.

4) Why do current solutions fail short?

The NLP community and the SW community are relatively isolated from each other, and often not aware of developments in the other community. For example, a recent discussion on an NLP mailing list showed that occasionally RDF (as an abstract data model) is confused with RDF/XML (as one RDF linearization) and rejected because of the verbosity of this linearization, even though other, more compact and more readable linearizations exist.

5) What insights do we need in order to reach a principled solution? What could a principled solution look like?

Co-located and/or interdisciplinary events. (Simply continue and extend the series of Multilingual Semantic Web and OntoLex workshops.) Interdisciplinary community groups.

6) How can standardization (e.g. by the W3C) contribute?

Standardization is actually a key issue here. The NLP community developed its own standards within the ISO, and succeeded in integrating different groups from NLP/computational linguistics/computational lexicography. Semantic Web standards, however, are standardized by the W3C. Even though, say, GrAF and RDF (see appendix) are conceptually very close, the potential synergies have been realized only recently. If these standardization initiatives could be brought in closer contact with each other, natural convergence effects are to be expected.

Appendix

Possible Future Convergences between Semantic Web and NLP

From the perspective of Natural Language Processing and Computational Linguistics, one of the developments I would expect for the next 5-10 years is the accelerating convergence of both disciplines, at least in certain aspects. On the one hand, this includes adopting Linked Data as a representation formalism for linguistic resources in; on the other hand, this includes the improved integration of NLP tools and pipelines in Semantic Web applications. Both developments can be expected to continue for the next decade.

The Prospective Role of Linked Data in Linguistics and NLP

In the last 20 years, Natural Language Processing (NLP) has seen a remarkable maturation, evident, for example, from the shift of focus of shared tasks from elementary linguistic analyses over semantic analyses to higher levels of linguistic description.[4] To a large extent, this development was driven by the increased adaption of statistical approaches during the 1990s. One necessary precondition for this development was the availability of large-scale corpora, annotated for the phenomena under discussion, and for the development of NLP tools for higher levels of description (say, semantics or anaphoric annotation), the number and diversity of annotations available (and necessary) increased continually.

During the same period, corpus linguistics has developed into a major line of research in linguistics, partially supported by the so-called “pragmatic shift” in theoretical linguistics, when scholars have recognized the relevance of contextual factors. The study of these context factors favored the application of corpora in linguistics at a broader scale, which can now be considered to be an established research paradigm in linguistics.

Taken together, both communities created increasingly diverse and increasingly large amounts of data whose processing and integration, however, posed an **interoperability challenge**. In a response to this, the NLP community developed generic formalisms to represent linguistic annotations, lexicons and terminology, namely in the context of the ISO TC37. As far as corpora are concerned, a standard, GrAF (Ide and Suderman, 2007), has been published this year. So far, GrAF is poorly supported with infrastructure and maintained by a relatively small community. However, its future application can take benefit of developments in the Linked Data community, where RDF provides a data model that is similar in philosophy and genericity, but that comes with a rich technological ecosystem, including data base implementations and query languages – which are currently not available for GrAF. Representing corpora in RDF, e.g., using an RDF representation of GrAF (Cassidy, 2010; Chiarcos, 2012), yields a number of additional benefits, including the uncomplicated integration of corpus data with other RDF resources, including lexical-semantic resources (e.g., WordNet) and terminology resources (e.g., GOLD). A comparable level of integration of NLP resources within a uniform formalism has not been achieved before, and to an increasing extent, this potential is recognized by researchers in NLP and linguistics, as manifested, for example, in the recent development of a Linguistic Linked Open Data cloud.[5] I expect this development, the adaption

[4] CoNLL Shared Tasks: 1999-2003 flat annotations (NP bracketing, chunking, clause identification, named entity recognition), 2004-2009: dependency parsing and semantic role labelling, 2010-2012: pragmatics and discourse (hedge detection, coreference).

[5] http://linguistics.okfn.org/llod

The importance of Semantic User Profiles and Multilingual Linked Data

Ernesto William De Luca
University of Applied Sciences Potsdam
Friedrich-Ebert-Strasse 4, 14467 Potsdam, Germany
deluca@fh-potsdam.de

1 Introduction

Today, people start to use more and more different web applications. They manage their bookmarks in social bookmarking systems, communicate with friends on Facebook[1] and use services like Twitter[2] to express personal opinions and interests. Thereby, they generate and distribute personal and social information like interests, preferences and goals [11]. This distributed and heterogeneous corpus of user information, stored in the user model (UM) of each application, is a valuable source of knowledge for adaptive systems like information filtering services. These systems can utilize such knowledge for personalizing search results, recommend products or adapting the user interface to user preferences. Adaptive systems are highly needed, because the amount of information available on the Web is increasing constantly, requiring more and more effort to be adequately managed by the users. Therefore, these systems need more and more information about users interests, preferences, needs and goals and as precise as possible. However, this personal and social information stored in the distributed UMs usually exists in different languages (language heterogeneity) due to the fact that we communicate with friends all over the world. Therefore, we believe that the integration of multilingual resources into a user model aggregation process to enable the aggregation of information in different languages which leads to better user models and thus to better adaptive systems.

1.1 The Use of Multilingual Linked Data

Because the Web is evolving from a global information space of linked documents to one where both documents and data are linked, we agree that a set of best practices for publishing and connecting structured data on the Web known as Linked Data. The Linked Open Data (LOD) project [4] is bootstrapping the Web of Data by converting into RDF and publishing existing available ”open datasets”. In addition, LOD datasets often contain natural language texts, which are important to link and explore data not only in a broad LOD cloud vision, but also in localized applications within large organizations that make use of linked data [2], [10].

The combination of natural language processing and semantic web techniques has become important, in order to exploit lexical resources directly represented as linked data. One of the major examples is the WordNet RDF dataset [13], which provides concepts (called *synsets*), each representing the sense of a set of synonymous words [7]. It has a low level of concept linking, because synsets are linked mostly by means of taxonomic relations, while LOD data are mostly linked by means of domain relations, such as parts of things, ways of participating in events or socially interacting, topics of documents, temporal and spatial references, etc. [10].

An example of interlinking lexical resources like EuroWordNet [14] or FrameNet[3] [1] to the LOD Cloud is given in [5] and [8]. Both create a LOD dataset that provides new possibilities to the lexical grounding of semantic knowledge, and boosts the “lexical linked data” section of LOD, by linking e.g. EuroWordNet and FrameNet to other LOD datasets such as WordNet RDF [13]. This kind of resources open up new possibilities to overcome the problem of language heterogeneity in different user models and thus allows a better user model aggregation [6].

2 Requirements for a User-Oriented Multilingual Semantic Web

Based on the idea presented above, some requirements have to be fulfilled:

Requirement 1: Ontology-based profile aggregation. We need an approach to aggregate information that is both application independent and application overarching. This requires a solution that allows to semantically define relations and coherences between different attributes of different UMs. The linked attributes must be easily accessible by applications such as recommender and information retrieval systems. In addition, similarity must be expressed in these defined relations.

Requirement 2: Integrating semantic knowledge. A solution to handle the multilingual information for enriching user profiles is needed. Hence, methods that incorporate information from semantic data sources such as EuroWordNet and to aggregate complete profile information have to be developed.

2.1 Multilingual Ontology-based Aggregation

For the aggregation of user models, the information in the different user models has to be linked to the multilingual information (as *Multilingual Linked Data*) as we want to leverage this information and use it for a more precise and qualitatively better user modeling. These resources can be treated as a huge semantic profile that can be used to aggregate user models based on multilingual information.

Figure 1 describes the general idea. The goal is to create one big semantic user profile, containing all information from the three profiles of the user information, were the

[1] http://www.facebook.com/

[2] http://twitter.com/

[3] http://framenet.icsi.berkeley.edu/

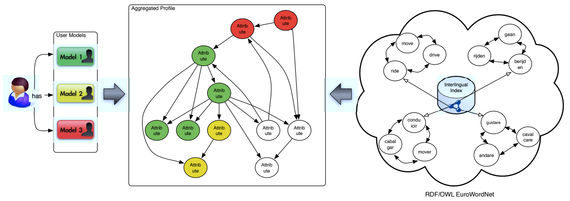


Figure 1: Integrating semantic knowledge about multilingual dependencies with the information stored in the user models.

data is connected. The first step is to add the multilingual information to the data contained in the different user models. This gives us a first model were the same data is linked together through the multilingual information.

2.2 Integrating Semantic Knowledge

The second step is to add links between data that is not linked through the multilingual information. The target is to have a semantic user model were data is not only connected on a language level but also on a more semantic similarity level. The aggregation of information into semantic user models can be performed similarly to the approach described in [3], by using components that mediate between the different models and using recommendation frameworks that support semantic link prediction like [12]. The combined user model should be stored in an commonly accepted ontology, like [9] to be able to share the information with different applications.

With such a semantic user model, overcoming language barriers, adaptive systems have more information about the user and can use this data to adapt better to the user preferences.

2.3 Conclusions

Analyzing the problems described above, we believe that more context information about users is needed, enabling a context sensitive weighting of the information used for the profile enrichment. The increasing popularity of Social Semantic Web approaches and standards like FOAF[4] can be one important step in this direction. On the other hand, multilingual semantic datasets itself (as for example multilingual linked data) have to be enriched with more meta-information about the data. General quality and significance information, like prominence nodes and weighted relations, can improve semantic algorithms to better compute the importance of paths between nodes. Enriching the quality of user profiles and the multilingual semantic representation of data can be helpful, because both sides cover different needs required for an enhancement and consolidation of a multilingual semantic web.

References

[1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA, 1998.

[2] Claudio Baldassarre, Enrico Daga, Aldo Gangemi, Alfio Massimiliano Gliozzo, Alberto Salvati, and Gianluca Troiani. Semantic scout: Making sense of organizational knowledge. In *EKAW*, pages 272–286, 2010.

[3] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286, 2008.

[4] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[5] Ernesto William De Luca. Aggregation and maintenance of multilingual linked data. In *Semi-Automatic Ontology Development: Processes and Resources*, pages 201–225. IGI Global, 2012.

[6] Ernesto William De Luca, Till Plumbaum, Jérôme Kunegis, and Albayrak. Multilingual Ontology-based User Profile Enrichment. In *Proc. Workshop on the Multilingual Semantic Web*, pages 41–42, 2010.

[7] A. Gangemi, R. Navigli, and P. Velardi. The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In R. Meersman and Z. Tari, editors, *Proc. of On the Move to Meaningful Internet Systems (OTM2003) (Catania, Italy)*, pages 820–838. Springer-Verlag, 2003.

[8] A. Gangemi and V. Presutti. Towards a Pattern Science for the Semantic Web. *Semantic Web*, 1(1-2):61–68, 2010.

[9] Dominik Heckmann, Tim Schwartz, Boris Brandherm, Michael Schmitz, and Margeritta von Wilamowitz-Moellendorff. Gumo - the general user model ontology. In *User Modeling*, pages 428–432, 2005.

[10] Andrea Nuzzolese, Aldo Gangemi, and Valentina Presutti. Gathering lexical linked data and knowledge patterns from framenet. In *The Sixth International Conference on Knowledge Capture (K-CAP 2011)in cooperation with the AAAI*, 2011.

[11] Till Plumbaum, Songxuan Wu, Ernesto William De Luca, and Sahin Albayrak. User modeling for the social semantic web. In *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, in conjunction with ISWC 2011*, 2011.

[12] Alexandrin Popescul and Lyle H. Ungar. Statistical relational learning for link prediction. In *Proceedings of the Workshop on Learning Statistical Models from Relational Data*, 2003.

[13] G. Schreiber, M. van Assem, and A. Gangemi. RDF/OWL Representation of WordNet. W3C Working Draft, W3C, June 2006. http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/.

[14] Piek Vossen. EuroWordNet General Document, Version 3, Final, 1999.

[4] http://www.foaf-project.org/

Thierry Declerck, DFKI
Abstract for Dagstuhl Seminar:The Multilingual Semantic Web (MSW)

1) What are in your view the most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web?
Answer: There is a (correct) statement that most knowledge is conveyed by Human Language, and therefore a criticism that you find in the semantic web (I consider here mainly the LOD/LD instantiation of the semantic web) only structured abstract knowledge representation. As a response to this criticism, our work can stress that language processing has to structure language data too, and that one of our task would be to represent structured language data in the same way as the knowledge objects, and to interlink those in a more efficient way as this has been done in the past, like for example in the simple/parole or the generative lexicon, linking thus language data "in use" with knowledge data "in use".

2) Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?
Answer: The possible approach sketched under point 1) whould be deployed in a multilingual fashion. If multilingual data is succesfully attached to knowledge objects, then multilingual and cross-lingual retrieval of knowledge is getting feasible. Not on the base of machine translation (only), but rather on the base of multilingual equivalents found linked to knowledge objects.
At the end not only knowledge of the world can be retrieved, but also knowledge of the words (or language) associated with the knowledge of the world. The knowledge of the language would be partial (no full grammar is to be expected),. But it can serve in many applications.

3) Which figures are suited to quantify the magnitude or severity of the problem?
Answer: I can not answer concretely this question. I also do not know if there is a real "problem". We could go on the way we are doing by now (searching Google or the like, using domain specific repositories, using Question/Answering systems for accessing knowledge in text, etc), but I expect a gain of efficiency in many natural language based application, dealing with the treatment of knowledge: semantic annotation, semantic disambiguation, information extraction, summarization, all in multi- and cross-lingual contexts. Terminology should also benefit from this approach (linking multilingual linguistic linked data with linked data), in offering a better harmonization of the domain specific terms used in various languages, while refering to establised terms used in the LD/LOD.

4) Why do current solutions fail short?
Answer: Well: all the natural language expressions available in knowledge objects are not (yet) available in a structured form, reflecting the knowledge of language. So that the linking of conceptual knowledge and language is done on a non-equilibrated manner: structured data on the one side and unalysed strings on the other one.

5) What insights do we need in order to reach a principled solution? What could a principled solution look like?
Answer: See the comment under point 1)

6) How can standardization (e.g. by the W3C) contribute?
Answer: Giving an consensual view on representation of the various types of knowledge and ways to integrate those, by merging (OWL?) or by mapping/linking (SKOS, lemon-LMF)

My possible conrtibution to the Workshop: Describing the potential LabelNet that can be resulting on the generalisation of linking structured language knowledge with domain knowledge. Generalizing the use of certain words/expressions (phrases, clauses, etc) so that labels (or linguistically described terms) can be re-used in different knowledge contexts.

There is also a specific domain I am working on, besides finance (XBRL; MFO) and radiology (RADLEX): Digital Humanities, more specifically two classification systems for tales and related genres. I am using there the Thompsom Motif Index and the Aarne Thompson Uther Type index of tales. Transformed those in explicit taxonomies. And we are currently working on representing the labels of such taxxonomies in LMF/lemon.

I could present actual work in any of these 3 domains, if wished.

Abstract for the Dagstuhl Multilingual Semantic Web Seminar

Bo Fu
{bofu}@uvic.ca
University of Victoria, British Columbia, Canada

In relation to realising cross-lingual data access on the multilingual semantic web, particularly through the use of mappings, a lack of collaboration support in the current research field appears to be an important problem that is yet to be addressed.

One of the best examples of collaboration on the web during the past decade is Wikipedia, which has successfully demonstrated the value and importance of collaboratively building domain knowledge in a wide range of subject matters. Similarly, on the semantic web, knowledge bases (i.e. ontologies and other formal specifications) regardless of their representations or syntaxes, are the wisdom of communities and likely to involve the effort of individuals and groups from many different backgrounds. Given these characteristics, it is thus important to provide the necessary support for collaborations that are taking place during various stages on the semantic web.

In recent years, we have seen ontology editors integrating collaboration features. For instance, WebProtégé [Tudorache et al., 2008] is designed to support the collaborative ontology editing process by providing an online environment for users to edit, discuss and annotate ontologies. This trend in providing collaboration support is not yet evident in other semantic web research fields. For example, research in ontology mapping generation and evaluation has focused on developing and improving algorithms to date, where little attention has been placed on supporting collaborative creation and evaluation of mappings.

Proceeding forward, one of the challenges for the multilingual semantic web is to design and develop collaboration features for tools and services, in order for them to

- support social interactions around the data[1], so that a group of collaborators working on the same dataset can provide commentary and discuss relevant implications on common ground;
- engage a wider audience and provide support for users to share and publish their findings, so that information is appropriately distributed for group decision making;
- support long-term use by people with distinct backgrounds and different goals, so that personal preferences can be fully elaborated; and
- enhance decision making by providing collaborative support from the beginning of the design process, so that collaborative features are included in the design process of tools and services to prevent these features being developed as an afterthought.

Reference

Tudorache T, Vendetti J, Noy N. (2008), *WebProtégé: A Lightweight OWL Ontology Editor for the Web*. In Proceedings of the 5th International Workshop on OWL: Experiences and Directions.

[1] In this context, *data* can be any variable related to applications on the semantic web. For example, it can be the results from ontology localisation, ontology mapping or the evaluations of such results.

Cross-lingual ontology matching as a challenge for the Multilingual Semantic Web

Jorge Gracia
Ontology Engineering Group
Universidad Politécnica de Madrid

Motivation

Recently, the Semantic Web has experienced significant advancements in standards and techniques, as well as in the amount of semantic information available online. Nevertheless, mechanisms are still needed to automatically reconcile information when it is expressed in different natural languages on the Web of Data, in order to improve the access to semantic information across language barriers. In this context several challenges arise [1], such as: (i) ontology translation/localization, (ii) cross-lingual ontology mappings, (iii) representation of multilingual lexical information, and (iv) cross-lingual access and querying of linked data.

In the following we will focus on the second challenge, which is *the necessity of establishing, representing and storing cross-lingual links among semantic information on the Web*. In fact, in a "truly" multilingual Semantic Web, semantic data with lexical representations in one natural language would be mapped to equivalent or related information in other languages, thus making navigation across multilingual information possible for software agents.

Dimensions of the problem

The issue of cross-lingual ontology matching can be explored across several dimensions

1. Cross-lingual mappings can be established at *different knowledge representation levels*, each of them requiring their own mapping discovery/representation methods and techniques:
   i. conceptual level (links are established between ontology entities at the schema level),
   ii. instance level (links are established between data underlying ontologies), and
   iii. linguistic level (links are established between lexical representations of ontology concepts or instances).
2. Cross-lingual mappings can be discovered *runtime/offline*. Owing to the growing size and dynamic nature of the Web, it is unrealistic to conceive a Semantic Web in which all possible cross-lingual mappings are established beforehand. Thus, scalable techniques to dynamically discover cross-lingual mappings on demand of semantic applications have to be investigated. Nevertheless, one can imagine some application scenarios (in restricted domains for a restricted number of languages) in which computation and storage of mappings for later reuse is a viable option. In that case, suitable ways of storing and representing cross-lingual mappings become crucial. Also mappings computed runtime could be stored and made available online, thus configuring a sort of pool of cross-lingual mappings that grows with time. Such online mappings should follow the linked data principles to favour their later access and reuse by other applications.

3. Cross-lingual links can be discovered either by *projecting* the lexical content of the mapped ontologies into a *common language* (either one of the languages of the aligned ontologies or a pivot language) e.g., using machine translation, or by *comparing the different languages directly* by means of cross-lingual semantic measures (e.g., cross-lingual explicit semantic analysis [2]). Both avenues have to be explored, compared, and possibly combined.

What is needed?

In summary, research has to be done in different aspects:

- *Cross-lingual ontology matching*. Current ontology matching techniques could be extended with multilingual capabilities, and novel techniques should be investigated as well.
- *Multilingual semantic measures*. Such novel cross-lingual ontology matching techniques above mentioned have to be grounded on measures capable of evaluating similarity or relatedness between (ontology) entities documented in different natural languages.
- *Scalability of matching techniques*. Although the scalability requirement is not inherent to the multilingual dimension in ontology matching, multilingualism exacerbates the problem due to the introduction of a higher heterogeneity degree and the possible explosion of compared language pairs.
- *Cross-lingual mapping representation*. Do current techniques for representing lexical content and ontology alignments suffice to cover multilingualism? Novel ontology lexica representation models [3] have to be explored for this task.

References

[1] J. Gracia, E. M. Ponsoda, P. Cimiano, A. G. Pérez, P. Buitelaar, and J. McCrae, "Challenges for the multilingual Web of Data," Journal of Web Semantics, vol. 11, pp. 63-71, Mar. 2012. Available at http://oa.upm.es/8848/1/Multiling.pdf

[2] P. Sorg and P. Cimiano, "Exploiting wikipedia for cross-lingual and multilingual information retrieval," Data & Knowledge Engineering, vol. 74, pp. 26-45, Apr. 2012. Available at http://dx.doi.org/10.1016/j.datak.2012.02.003

[3] J. McCrae, G. A. de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner, "Interchanging lexical resources on the semantic web," Language Resources and Evaluation, vol. 46, 2012. Available at http://dx.doi.org/10.1007/s10579-012-9182-3

Multilingual Access to the Web: the UKP Lab’s Perspective

Iryna Gurevych, http://www.ukp.tu-darmstadt.de/

Ubiquitous Knowledge Processing (UKP Lab), Technische Universität Darmstadt

We first outline a set of research directions for the multilingual content processing on the web, such as aggregating the knowledge in multiple documents, assessing the quality of information, engineering complex multilingual Web-based systems, and scalability of machine learning based approaches to new tasks and domains. Then, we present some research initiatives at UKP Lab with immediate relevance to the research directions listed above.

Research directions

The volume of text-based data, especially user-generated content in many languages, on the Web has been continuously growing. Typically, there are multiple documents of various origins describing individual facets of the same event. This entails redundancy, resulting in the need to **aggregate the knowledge distributed across multiple documents**. It involves the tasks such as removing redundancy, information extraction, information fusion and text summarization. Thereby, the intention of the user and the current interaction context play an important role.

Another fundamental issue in the Web is **assessing the quality of information**. The vast portion of the content is user-generated and is thus not subject to editorial control. Therefore, judging its quality and credibility is an essential task. In this area, text classification methods have been applied and combined with social media analysis. Since the information on the Web might quickly become outdated, advanced inference techniques should be put to use in order to detect outdated content and controversial statements found in the documents.

Due to advances in ubiquitous computing and the penetration of small computer devices in everyday life, the integration of multiple knowledge processing techniques operating across different modalities and different languages on huge amounts of data has become an important issue. This is an issue with **challenges to be addressed in software engineering**. It requires standardization of the interface specifications regarding individual components, ensuring the scalability of approaches to large volumes of data, large user populations and real-time processing, and solutions regarding the technical integration of multiple components into complex systems.

Current multilingual language processing systems extensively utilize machine learning. However, the training data is lacking in many tasks and domains. To alleviate this problem, the **use of semi-supervised and unsupervised techniques** is an important research direction. For the supervised settings, utilizing **crowdsourcing and human computation** such as Amazon Mechanical Turk, Games with a Purpose, or Wiki-based platforms for knowledge acquisition is a current research direction [1]. Research is needed to find ways of efficiently acquiring the needed high-quality training data under the time and budget constraints depending on the properties of the task.

Research Initiatives at UKP Lab

The above research directions have been addressed in several projects by UKP Lab at the Technische Universität Darmstadt, described below.

Sense-linked lexical-semantic resources. We present a freely available standardized large-scale lexical-semantic resource for multiple languages called **UBY**[1] [3,5]. UBY currently combines collaboratively constructed and expert-constructed resources for English and German. It is modeled according to the ISO standard Lexical Markup Framework (LMF). UBY contains standardized versions of WordNet, GermaNet, FrameNet, VerbNet, Wikipedia, Wiktionary and OmegaWiki. A subset of the resources in UBY is linked at the word sense level, yielding so-called mono- and cross-lingual sense alignments between resources [6,7,8]. The UBY database can be accessed by means of a Java-based API available at Google Code (http://code.google.com/p/uby) and used for knowledge-rich language processing, such as word sense disambiguation.

Multilingual processing based on the Unstructured Information Management Architecture (UIMA). We put a strong focus on component-based language processing (NLP) systems. The resulting body of software is called the **Darmstadt Knowledge Processing Software Repository** (DKPro) [2]. Parts of DKPro have already been released to the public as open source products, e.g.:

- **DKPro Core**[2] is an integration framework for basic linguistic preprocessing. It wraps a number of NLP tools and makes them usable via a common API based on the Apache UIMA framework. From the user perspective, the aim of DKPro Core is to provide a set of components that work off-the-shelf, but it also provides parameter setting options for the wrapped tools. The roadmap for DKPro Core includes: packing models for the different tools (parser, tagger, etc.) so they can be logically addressed by name and version and downloaded automatically, cover more tagsets and languages, logically address corpora and resources by name and version and download them automatically, provide transparent access to the Hadoop HDFS so that experiments can be deployed on a Hadoop Cluster.
- **DKPro Lab**[3] is a framework to model parameter-sweeping experiments as well as experiments that require complex setups which cannot be modeled as a single UIMA pipeline. The framework is lightweight, provides support for declaratively setting up experiments, and integrates seamlessly with Java-based development environments. To reduce the computational effort of running an experiment with many different parameter settings, the framework uses dataflow dependency information to maintain and reuse intermediate results. DKPro Lab structures the experimental setup with three main goals: facilitating reproducible experiments, structuring experiments for better understandability, structuring experiments into a workflow that can potentially be mapped to a cluster environment. In particular, the latter is currently in the focus of our attention.

The DKPro software collection has been employed in many NLP projects. It yielded excellent performance in a series of recent language processing shared tasks and evaluations, such as:

(i) **Wikipedia Quality Flaw Prediction Task** in the PAN Lab at CLEF 2012. [9]

(ii) **Semantic Textual Similarity Task** for SemEval-2012, held at *SEM (the First Joint Conference on Lexical and Computational Semantics). [4]

(iii) **Cross-lingual Link Discovery Task** (CrossLink) at the 9th NTCIR Workshop (NTCIR-9), Japan. [10]

[1] http://www.ukp.tu-darmstadt.de/uby

[2] http://code.google.com/p/dkpro-core-asl/

[3] http://code.google.com/p/dkpro-lab/

Literature

1. Iryna Gurevych and Torsten Zesch. (Editors). 2012. **Special issue on "Collaboratively Constructed Language Resources" in the Language Resources and Evaluation Journal**. Published online http://www.springerlink.com/content/c22w62t3784j0651/?MUD=MP, printed in fall 2012. (To appear.)
2. Richard Eckart de Castilho and Iryna Gurevych. **A Lightweight Framework for Reproducible Parameter Sweeping in Information Retrieval**. In: *Proceedings of CIKM 2011 Workshop DESIRE: Data infrastructures for Supporting Information Retrieval Evaluation Proceedings*, Glasgow, UK, September 2011, pp. 7-10.
3. Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek and Christian M. Meyer. **UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF**, In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pp. 275-282, ISBN 978-2-9517408-7-7, May 2012.
4. Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. **UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures.** In: Proceedings of the 6th International Workshop on Semantic Evaluation, in conjunction with the 1st Joint Conference on Lexical and Computational Semantics, pp. 435-440, June 2012.
5. Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer and Christian Wirth. **Uby - A Large-Scale Unified Lexical-Semantic Resource Based on LMF**, In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), pp. 580-590, April 2012.
6. Judith Eckle-Kohler and Iryna Gurevych. **Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability**, In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), pp. 550-560, April 2012.
7. Christian M. Meyer and Iryna Gurevych. **What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage**, In: *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thailand, pp. 883-892, September 2011.
8. Elisabeth Niemann (geb. Wolf) and Iryna Gurevych. **The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet**. In: *Proceedings of the International Conference on Computational Semantics (IWCS)*, Oxford, UK, pp. 205-214, January 2011.
9. Oliver Ferschke, Iryna Gurevych, Marc Rittberger. (2012) **FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia**. In: *Proceedings of PAN 2012 Lab "Uncovering Plagiarism, Authorship, and Social Software Misuse" held in conjunction with the CLEF 2012 Conference* (Shared task: Quality Flaw Prediction in Wikipedia), 17-20 September 2012, Rome, Italy. (To appear.)
10. Jungi Kim and Iryna Gurevych. **UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery,** In: *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pp. 487-494, December 2011.

Overcoming Linguistic Barriers
to the Multilingual Semantic Web

Graeme Hirst
Department of Computer Science, University of Toronto
gh@cs.toronto.edu

29 June 2012

Sometime between the publication of the original Semantic Web paper by Berners-Lee, Hendler, and Lassila (BLHL) (2001) and Berners-Lee's (2009) "Linked Data" talk at TED, the vision of the Semantic Web contracted considerably. Originally, the vision was about "information"; now it is only about data. The difference is fundamental. Data has an inherent semantic structure and an *a priori* interpretation. Other kinds of information need not. In particular, information in linguistic form gains an interpretation only in context, and only for a specific reader or community of readers.

I do not mean to criticize the idea of restricting our Semantic Web efforts to data *pro tem*. It is still an extremely challenging problem, and the results will still be of enormous utility. At the same time, however, we need to keep sight of the broader goal, and we need to make sure that our efforts to solve the smaller problem are not just climbing trees to reach the moon.

In the original vision, "information is given well-defined meaning" (p. 37), implying that it didn't have "well-defined meaning" already. Of course, the phrase "well-defined meaning" lacks well-defined meaning, but BLHL are not saying that information on the non-Semantic Web is meaningless; rather what they want is precision and lack of ambiguity in the Semantic layer. In the case of linguistic information, this implies semantic interpretation into a symbolic knowledge representation language of the kind they talk about, which is a goal that exercised, and ultimately defeated, research in artificial intelligence and natural language understanding from the 1970s through to the mid-1990s.

One of the barriers that this earlier work ran into was the fact that traditional symbolic knowledge representations — the kind that we still see for the Semantic Web — proved to be poor representations for linguistic meaning, and hierarchical ontologies proved to be poor representations for the lexicon (Hirst 2009). Near-synonyms, for example, form clusters of related and overlapping meanings that do not admit a hierarchical differentiation. And quite apart from lexical issues, any system for representing linguistic information must have the expressive power of natural language; we have nothing anywhere close to this as yet.

All these problems are compounded when we add multilinguality as an element. For example, different languages will often present a different and mutually incompatible set of word senses, as each language lexicalizes somewhat different categorizations or perspectives of the world, and each language has *lexical gaps* relative to other languages and to the categories of a complete ontology. It is rare even for words that are regarded as *translation equivalents* to be completely identical in sense; more usually, they are merely cross-lingual near-synonyms (Hirst 2009).

And then we have the problem of querying linguistic information on the Semantic Web, again in a natural language. Much of the potential value of querying the Semantic Web is that the system may act on behalf of the user, finding relevance in, or connections between, texts that goes beyond anything the original authors of those texts intended. That is, it could take a *reader-based view of meaning*, "What does this text mean to me?" (Hirst 2008). The present construal of the Semantic Web, however, is limited to a *writer-based view* *of meaning*. That is, semantic mark-up is assumed to occur at page-creation time, either automatically or semi-automatically with the assistance of the author (Berners-Lee et al. 2001); a page has a single, fixed, semantic representation that (presumably) reflects its author's personal and linguistic worldview and which therefore does not necessarily connect well with queries to which the text is potentially relevant.

But that's not to say that author-based mark-up isn't valuable, as many kinds of natural language queries take the form of intelligence gathering, "What are they trying to tell me?" (Hirst 2008). Rather, we need to understand its limitations, just as we understand that the query "Did other people like this movie?" is an imperfect proxy for our real question, "Will *I* like this movie?"

This gives us a starting point for thinking about next steps for a monolingual or multilingual Semantic Web that includes linguistic information. We must accept that it will be limited, at least *pro tem*, to a static, writer-based view of meaning. Also, any semantic representation of text will be only partial, and will be concentrated on facets of the text for which a representation can be constructed that meets BLHL's criterion of relative precision and lack of ambiguity, and for which some relatively language-independent ontological grounding has been defined. Hence, the representation of a text may be incomplete, patchy, and heterogeneous, with different levels of analysis in different places (Hirst and Ryan 1992).

We need to recognize that computational linguistics and natural language processing have been enormously successful since giving up the goal of high-quality knowledge-based semantic interpretation 20 years ago. Imperfect methods based on statistics and machine learning frequently have great utility. Thus there needs to be space in the multilingual Semantic Web for these kinds of methods and the textual representations that they imply — for example, some kind of standardized lexical or ontolexical vector representation. We should expect to see symbolic representations of textual data increasingly pushed to one side as cross-lingual methods are developed in distributional semantics (Evert 2013) and semantic relatedness. These representations don't meet the "well-defined meaning" criterion of being overtly precise and unambiguous, and yet they are the representations most likely to be at the centre of the future multilingual Semantic Web.

References

Berners-Lee, Tim (2009). The Next Web. TED Conference, Long Beach, CA. `http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html`

Berners-Lee, Tim; Hendler, James; and Lassila, Ora (2001). The Semantic Web. *Scientific American*, 284(5), May 2001, 34–43.

Evert, Stefan (2013). *Distributional Semantics*. Morgan & Claypool (to appear).

Hirst, Graeme (2008). The future of text-meaning in computational linguistics. In: Sojka, Petr; Horák, Aleš; Kopeček, Ivan; and Pala, Karel (editors), *Proceedings, 11th International Conference on Text, Speech and Dialogue (TSD 2008)*, (Lecture Notes in Artificial Intelligence 5246), Berlin: Springer-Verlag, 2008, 1–9.

Hirst, Graeme (2009). Ontology and the lexicon. In: Staab, Steffen and Studer, Rudi (editors), *Handbook on Ontologies* (second edition), Berlin: Springer-Verlag (International Handbooks on Information Systems), 2009, 269–292.

Hirst, Graeme and Ryan, Mark (1992). "Mixed-depth representations for natural language text." In: Jacobs, Paul S. (editor). *Text-based Intelligent Systems*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1992, 59–82.

An Event-type Driven Ontology for Multilingual Semantic Web
A position statement for
The Dagstuhl Seminar on Multilingual Semantic Web

Chu-Ren Huang, The Hong Kong Polytechnic University

1) The most crucial challenge to a multilingual semantic web is its accessibility and inter-operability for people from different linguistic and cultural backgrounds, a challenge that the currently envisioned shared ontology could compound rather than ameliorate. First, the semantic web and its content must be accessible to people from different parts of the world using different languages and having different culturally conventionalized world views. This issue requires both multilingual language resources and culture-specific ontology, presumably linked through and mapped to a shared ontology. Second, I consider the inter-operability issue crucial but from a maverick perspective. It is crucial to recognize that for all the tasks performed on the semantic web, each of them comes with an intentional goal with culture-specific background. Taking Tim Berners-Lee's example of buying flowers, the typical and most likely event type in the West is to buy loose flowers or bouquets for someone dear to the buyer; but in the Chinese context, the typical event is to buy flower basket installations for social networking functions such as a wedding or opening of a business. Similarly, when searching for a family dinner out, a diner or a "family-oriented" restaurant (such as Appleby's) with crayons for children are typical for users in U.S. But in the Chinese context, a Chinese meal with round-table seating may be crucial. I view inter-operability challenge as 1) to be able to identify these common event types as well as their event structure skeletons and cultural variants for integration of information; and 2) to allow culture/domain specific event-driven tasks to exploit knowledge content encoded in either the shared ontology or a domain specific ontology. It is of course important to note that these event-variation issues are often embedded in language and need to be described in languages accessible to the users.
2) Accessibility and Inter-operability as described above is critical to whether an industry based on SW can deliver or not. In the multilingual and multicultural world connected by the SW, localization will not be as effective as SW is concept-based, not text-based; and in our increasingly multi-cultural world, a user's assumed cultural convention is rarely simple and very often not determined by the language s/he uses (or his/her IP).

   All sectors should be affected. However, this challenge should be particularly keen for the following sectors, (1) creative culture industries (CCI), including but not limited to (culture) tourism, hospitality, (digital) museums, etc., (2) health communication and health care information providers, (3) advertisement

[Second point (2) is skipped as no relevant data can be given, underlining how difficult it is to characterize the cultural/linguistic background of web users.]

3) It is likely that the current solutions fail short because they focus on ensuring the meaning content is accessible to different machines but not to how the information can be utilized or interpretable by human users in the world. It is also important to note that

-there are no (or at least very rare) large scale culturally sensitive knowledgebases.

-construction of ontologies, including domain-specific ontologies, so far focused on shared, not differential knowledge structure.

4) People often act on both personal experience and culturally conventionalized shared experience. Note these experiences can correspond to shared knowledge, but do not necessarily follow the knowledge structure represented in an upper ontology. In addition, such behaviors are driven by the goal of the person (i.e. the telic cause of Aristotle).

Take the treatment of environmental and ecology issues for example. It seems that all issues with environmental impact can be boiled down to those which impacts *feeding* or *breeding* for the living organisms involved. However, feeding for person, for a cow, or for a microbe, involves very different participants and very different environmental conditions; the current approach to SW ontologies seems to require that these events are given radically different representations. Another good example involves emotions. Although there are culture- and species-specific variations to expressing and reaction to emotions, it is generally accepted that people recognize the same emotion types across different cultures: anger, fear, etc. The recognition of these common event types (e.g. *feeding*, *breeding, fear, happiness,* etc.) given different contextual information will endow SW with powerful and effective means to deliver what users really want.

The solution is an additional dimension of ontology based on event types; in addition to the current entity type based shared ontology. The Chinese writing system as an ontology based the basic concept represented by the radicals is a good example as I have shown previously that the conceptual cluster sharing a radical is often based on event-relations such as telic and agentive, and less often by entity-type relations such as is-a or part-of.

5) Standardization should help provided that we do thorough study to explore the common event-types which are crucial to human activities and draft event-driven ontologies based on the research. An especially difficult challenge is to capture

the telic event types, being able to link telic goals to events that are necessary to attend that goal will allow SW to work on both meaning stated and intension/need expressed.

Dagstuhl Seminar on the Multilingual Semantic Web
Abstract

Nancy Ide, Department of Computer Science,
Vassar College, Poughkeepsie, New York, USA

July 31, 2012

A "language- and culture-neutral" Semantic Web (SW) will have to accommodate different linguistic and cultural perspectives on the knowledge it contains, in the same way as it must accommodate temporal, spatial, etc. perspectives. Ins the long term, probably the greatest challenge for SW development is to seamlessly handle a multiplicity of viewpoints (including language) on knowledge.

In the short term, we can do what we can with what we have now. One hypothesis is that a multilingual SW can be achieved–or at least approached–by mapping other languages to the broad range of existing ontological vocabularies that have been developed, almost exclusively in English, for various topics and domains. Existing language resources that cover multiple languages–notably, resources such as WordNet and FrameNet, but also bi- and multi-lingual lexicons developed in, for example, large EU projects over the past two decades–could be exploited for this purpose.

A major step toward a multilingual SW, and toward interoperability for mono- and multi-lingual language resources in general, would be to render lexical and other language resources as linked data (as WordNet and FrameNet already are). As linked data, the resources will have achieved some degree of structural ("syntactic") interoperability (to the extent that the relations and properties used in their representations are defined consistent). A linked data representation will also make a move toward conceptual ("semantic") interoperability (see [5]), because the various resources can, in principle, be linked to each other, either directly or via mediator ontologies that provide an exchange reference model for linking between resources (see, for example, Chiarcos' Ontologies of Linguistic Annotation (OLiA) [1]). While mapping concepts among lexicons and similar resources is notably difficult, some immediate steps can be made by linking the linked lexicons with corpora annotated with the categories defined in them. For example, the Manually Annotated Sub-Corpus (MASC) [4] has been rendered in RDF, and its WordNet and FrameNet annotations have been linked to the relevant entries in the linked data versions of these resources

[2]. In this form, all three resources are automatically combined, and SPARQL queries can be used to access, for example, all occurrences annotated with a specific FrameNet frame element, or all words used in a particular WordNet sense. Additionally, one could query to find the FrameNet frames that correspond to a particular WordNet sense via the corpus annotation, thus providing input for a better conceptual mapping of the two resources themselves. Ultimately, any annotated word, phrase, entity, etc. could be linked to occurrences of the same phenomenon in data in other languages, either directly or via an interlingua (if appropriate). The potential of such massively interlinked multilingual data for NLP research and development is enormous.

Rendering language resources as linked data so that they can be used together seamlessly requires a consistent model of the phenomena in question, including not only ontological concepts but also their inter-relations and properties. There have been efforts to devise standards which, although not specifically aimed toward linked data, provide underlying models of lexical data [3] that are isomorphic to the linked data model. However, there is much more (relatively mundane) work to be done in order to ensure compatibility in the domain model. As a simple example, consider modeling the general concept "Annotation" by defining its relations to other concepts and properties, with an eye toward enabling relevant queries. This requires answering (seemingly) simple questions like, does an Annotation have a single target, or can it have several? If it has several, does it apply to each individually (e.g., a "noun" annotation applied to all text spans identified as nouns throughout a text), or does it apply to an aggregation of its targets (e.g., a "verb" annotation applied to two discontiguous text spans that comprise a single verb)? How do we distinguish these two cases in the model? Etc. While this seems trivial on the one hand, the different communities for whom "annotation" is a relevant concept–including not only computational linguists but also humanities scholars who annotate in the more traditional sense, as well as annotators of images, video and audio, etc.–must adopt a common, or at least compatible, model if their data is to be used together (which much of it will ultimately be), and if we do not want to be faced with a different form of query for each case. This is where standards-making bodies, like W3C and ISO, must critically step in.

Groups such as the W3C and ISO can foster the development of a (multilingual) SW by:

- Promoting the use of SW technologies to structure and describe existing and future language resources, including lexicons, ontologies, and corpora developed by the NLP community;

- Establishing stronger ties with the NLP community, to leverage their expertise in identifying/extracting information;

- Continuing the top-down development of the SW infrastructure;

- Developing best practice guidelines for domain modeling in RDF/OWL;

- Overseeing and coordinating bottom-up development of RDF/OWL models for specific bodies of knowledge developed by specific disciplines/communities/interests;
- Actively seeking to identify commonalities among the varied models and bodies of knowledge and ensuring that efforts are combined/harmonized;
- Working on ways to accommodate different views of knowledge, including language and culture, in the SW.

Comprehensive interoperability via standardization is a long-term goal that is unlikely to be achieved anytime soon. This means that for the interim, we have to explore effective ways to bridge the differences in the concepts and structure of knowledge sources. Representing existing language resources as linked data is one way to approach that problem.

References

[1] Christian Chiarcos. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16, 2008.

[2] Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. Towards open data for linguistics: Linguistic linked data. In Iryna Gurevych and Jungi Kim, editors, *The Peoples Web Meets NLP: Collaboratively Constructed Language Resources*. Springer, forthcoming.

[3] Gil Francopoulo and Monty George et al. Lexical markup framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, 2006.

[4] Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. The Manually Annotated Sub-Corpus : A community resource for and by the people. In *Proceedings of the The 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010.

[5] Nancy Ide and James Pustejovsky. What does interoperability mean, anyway? Toward an operational definition of interoperability. In *ICGL 2010: Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong, China, 2010.

Multilingual Semantic Web - a personal perception

Antoine Isaac, Europeana

Caveat: the views here represent a personal take on MSW issues, resulting from involvement in Europeana.eu (providing access to 24 millions objects from 33 countries) and other projects in the cultural sector, as well as my experience with SKOS and ontology alignment. It shall be taken with a grain of salt: multilingual issues have not been the exclusive focus of my work and I may miss some efforts, especially recent ones. Which in this case tells about their visibility.

Technical & practices issues

SW technology does a lot to enable ML: tags on RDF literals allow for language-specific data, and the aggregation of those language-specific literals allow for "multilingual entities". Some ontologies enable easy representation of e.g., multilingual concepts (SKOS). There are even finer-grained models available (Lexinfo). But there are seldom known and used in large datasets.
As a result some technical issues still don't have commonly shared solution. For example, representing "translation of a statement". Consider the following:

```
ex:book dc:subject "multilingual semantic web"@en ;
        dc:subject "challenges"@en ;
        dc:subject "web sémantique multilingue"@fr .
```

This is a typical example of bibliographic data ported to RDF in a very basic way. It does not fully represent translation links and thus fails to some applications.
There should also be more attention given to patterns for giving the language of an object (e.g. a web page or a video recording) vs. the language of data about that object or the language of an interface. Technology (e.g. through the "one-to-one principle") makes clear what can be done; but in some domains (Europeana) data providers or consumers may still be confused.

Availability of tools and data

Many tools of reference in the SW community are mostly monolingual [Olensky]. A lot of datasets, and most experimentations and case studies are in English. More precisely, there are may resources available for a multilingual SW:
- terminological/conceptual bases (wordnets, SKOS datasets, dictionaries...)
- language processing tools (translators, language recognizers, parsers...)

Yet these are difficult to find. Few inventories exist, mostly paper reports not easily exploitable. We need a better way of gathering and sharing information on MSW resources. Here, specific metrics can be useful, for evaluating multilingual tools and measuring the "multilingual quality" of datasets. One starting point would be to indicate the language(s) covered by datasets on the linked data cloud, the labels per language, etc, refining for example language-related quality criteria used in the SKOS community (e.g., [Mader])
Further, many relevant resources are not open and/or are published in a format that does not enable easy re-use. This the case for many wordnets, and a true pity for resources that are created with public money. Some communities have made progress in releasing multilingual datasets, but a lot remains to be done.
In relation with the above metrics, we guidelines could help (if not standards) both to provide MSW resources or to select them for consumption.
To compensate the rarity of tools and resource, we should also be open to "less AI-focused" solutions, such as crowdsourcing translation or multilingual tagging.

Community organization and awareness

The above issues are partly caused by the homogeneity of the “core” SW community, mostly academic and English-speaking. That prevents diversity to emerge re. experiments and tools. It also makes it harder to be aware of relevant efforts in other communities (information retrieval, databases, more general web community); if just because these communities have the same bias...

Further, the difficulty of "traditional" problems is raised an order of magnitude higher when transferred from a monolingual context to a multilingual one (NB: that applies both re. finding and implementing solutions, and evaluating them). Bluntly put: working on a multilingual problem is not the most effective way of getting a paper published, and that does not help. For example the Ontology Alignment Evaluation Initiative [OAEI] has featured multilingual tracks for a while. Bar a few exceptions [Meilicke], participation has often been low.
Maybe the current SW community is not the ideal forum for tackling multilingual problems. Or it may just be able to progress on very specific issues (e.g. focusing on producing and sharing data). On evaluation matters, especially it could help to better share efforts (corpora, gold standards, methods and measures) with other communities—e.g., databases, web (services) or information retrieval. A relevant initiative is CLEF [CLEF].
Besides, there are many vendors that propose relevant solutions, esp. in "language technology". But they probably find it difficult to relate to the SW community. As long as vendors make a reasonable benefit in their current (non-SW) environment many won’t seriously move and liaise with academic efforts.
We need to getting more varied people interested and contributing to the MSW issues—but maybe from their own communities' perspectives.

Use cases
For bringing people together, it would help to identify the most relevant features of (end user) application scenarios. One usual example is localization: adapt a display depending on the language/country selected. This imposes multilingual requirements both on ontologies and instance data level. But there are other dimensions to multilingual access in which semantic web technology can be relevant: query translation, document translation, information extraction and data enrichment, browsing and personalization, knowledge acquisition for non English speakers, interaction between users and system... Some of these are neither strictly multilingual nor semantic web-specific. In such case, the potential added value of MSW should be detailed: for example, enhancing search results in one language based on links established using semantic resources in another language.
Maybe such a gathering focus more on cases where multilinguality is really crucial. For example, Europeana is encouraging application of SW technology for access to culture resources, where all EU languages should ultimately been tackled. It especially envisions tapping into a semantic data layer, which involves alignment of multilingual vocabularies and metadata enrichment. In a completely different domain, the VOICES project envisions using linked data technology for social and rural development. Key issues there are sharing locally produced data where local languages are more important than English, and building a robust data-to-speech service [De Boer].

References

[CLEF] Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum. http://www.clef-initiative.eu/

[De Boer] Victor de Boer, Pieter De Leenheer, Anna Bon, Nana Baah Gyan, Chris van Aart, Christophe Guèret, Wendelien Tuyp, Stephane Boyera, Mary Allen, Hans Akkermans. RadioMarché: Distributed Voice- and Web-interfaced Market Information Systems under Rural Conditions. International Conference on Advanced Information Systems Engineering, CAiSE'2012.

[Mader] Christian Mader, Bernhard Haslhofer, Antoine Isaac. Finding Quality Issues in SKOS Vocabularies. TPDL 2012, http://arxiv.org/abs/1206.1339v1

[Meilicke] Christian Meilicke, Raúl García Castro, Fred Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondrej Svab-Zamazal, Vojtech Svatek, Andrei Tamilin, Cássia Trojahn, Shenghui Wang. MultiFarm: A Benchmark for Multilingual Ontology Matching. Journal of Web Semantics, to appear.

[OAEI] http://oaei.ontologymatching.org/

[Olensky] Marlies Olensky. Market study on technical options for semantic feature extraction. Europeana v2.0 technical report, http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Market+study+on+technical+options+for+semantic+feature+extraction

Acknowledgements: feedback on this paper was provided by Vivien Petras, Juliane Stiller, Marlie Olensky (Humboldt University Berlin)

Christian Lieske, SAP AG

When people start to think about cultural diversity, they sooner or later start to talk about mouth-watering dishes. Thus, I have taken the freedom to choose a title for my position statement that alludes to a recipe (see below). As in almost any recipe, important details are missing from my recipe – the details can only be revealed face-to-face/in a joint session of practice. Accordingly, I would be happy to be invited to elaborate my position/my input at the seminar.

Rather than providing direct references, I provide pointers that indicate with which kind of background I tend to look at things.

Please note: All my thoughts are my own and not endorsed by my employer in any way.

10 Ingredients for the Multilingual Semantic Web Delight

1. World-ready Web Stack

Look for example at the mailing lists run by the W3C Internationalization Activity, or the Unicode Consortium to see that even mainstream topics such as HTML5, CSS3, JavaScript and Unicode are still undergoing modification in order to cover multilingual/-cultural dimensions more complete, or enhanced.

2. Concept-based Content Creation

Realize how model-based approaches (doesn't matter if your model entities are objects or events) already help to generate expressions in multiple idioms. In addition, read up for example on the Wikidata project to sense that there is a value in language-neutral representations.

3. Connected Organizational Constituencies

Be surprised that the problem space of that EC project epSOS is overlapping with the "Spanish patient needs medication in Germany" scenario that shows up in Semantic Web scenarios.

4. Transparency on Stakeholders/Contributors and Contribution Framework

Don't be blind to the fact that enterprises may have much to contribute to the Semantic Web. Don't you think for example that providers of pharmaceutical companies already have databases that capture relationships between the incarnations of their products on different markets?

5. Eye on Multi-Modality

Acknowledge that human interaction and information dissemination is not just based on written text. Consider how to take care of graphics/images or sound/voice (especially considering findings on the situation in the non-Western world from initivaties such as the World Wide Web Foundation or the Praekelt Foundation).

6. Anyone-Anytime-Anywhere Paradigm

Provide tooling that doesn't require a diploma in SPARQL, and the installation of a heavyweight application. If you want for example contributions to ontologies or vocabularies think "Point your mobile browser to a URL, and comment". Understand for example tools like translation memories engines that assist in language-related activities.

7. Reuse/Minimalism, and Clean, Open, Traceable Information Sources

Ask yourself how much trust you would have in a HTML table that would tell you "The drug that is called X in Spain is called Y Germany". Wouldn't you for example like to see provenance information before you order the drug?

8. Open-Minded NLP Community

Be aware of the fact that "mapping" is a very simple mathematical function. Do we think for example that a mapping will suffice to go from a blood pressure as measured in Germany to one that can be understood by a French speaking physician? Or does the "mapping" concept that is favoured by the NLP community need to be rethought?

9. Non-functional Requirements

Don't underestimate that you may need to prioritize, and schedule roadmap items. Otherwise, decision makers may not see the relevance and importance of Semantic Web activities.

10. Implementation-backed Standards and Best Practices

Makes sure that you have implementations for standards and best practices. Think for example how much easier the creation of multilingual Web sites would be if all Web server downloads would come with a "template" for multilanguage/multi-country Web presences.

== Background for Details ==

MultilingualWeb, Linked Open Data & EC "Connecting Europe Facility"
http://www.w3.org/International/multilingualweb/luxembourg/slides/d4-lieske.pdf

Meta Data for the Masses http://www.tekom.de/upload/3138/TERM12_Lieske.pdf

Best Practices and Standards for Improving Globalization-related Processes
http://www.w3.org/International/multilingualweb/pisa/slides/lieske.pdf

The Bricks to Build Tomorrow's Translation Technologies and Processes
http://www.w3.org/International/multilingualweb/madrid/slides/lieske.pdf

Whole World OLIF - Version 3.0 of the Versatile Language Data Format
http://www.tekom.de/upload/2284/OASIS_40_Lieske.pdf

Morphing into Mobile Multilinguality http://www.localizationworld.com/lwparis2012/E5.pptx

Content Quality Management with Open Source Language Technology
http://www.tekom.de/upload/3383/TA24_Lieske-Sasaki.pdf

Giving Prosoday a Meaning http://www.sics.se/~gamback/publications/eurospeech97.ps

Compositional Semantics http://researchr.org/publication/BosGLMPW96

Dagstuhl abstract: Moving beyond labels

John McCrae – CITEC, Universität Bielefeld

July 16, 2012

While some of the key resources in the Semantic Web, notably DBPedia, have placed considerable effort on internationalisation of their resources, most of the vocabularies including some of the widely used vocabularies, such as FOAF and even W3C standards such as RDFS, fail to provide labels in any language other than English. Even worse, many of these resources fail to even indicate that these labels are in English by means of standard meta-data. A clear issues with providing multilingual data as labels is that, all the emphasis on providing labels in languages other than English is on the data provider. It is of course extremely unlikely that for all but the largest of data providers, they could provide and check translations for even the Top 10 languages[1]. Even worse, for many applications that wish to use linked data and the Semantic Web, more lexical information is required than just a simple label. In particular, for many applications, such as question answering and natural language generation[2], information such as part-of-speech, morphology (e.g., irregular inflections) and syntactic information, such as subcategorization, would be extremely helpful. Anomalously, a simple and clear solution to this is available: by linking to dictionaries and lexica we can clearly define these concepts along with their multilingual equivalents.

A recent paper by Basil Ell *et al.*[1] recently claimed that only 0.7% of the entities in the web of data had labels in a language other than English, while 38.3% of the data had English labels. As such it is clear that the adoption of multilingual linked data within industry and research has been severely limited. Assuming that these organisations do not have the resources to provide translations for most language, a key issue is how these translation may be sourced from third parties. A solution to this may be to provide a central repository for localisation of vocabularies and data, i.e., a Google or Facebook of multilingual data. While this solution may have many clear advantages it seems unlikely that any existing service provider would step up to fill this rôle nor that it would a profitable new venture. As such, it seems that this solution is unlikely to materialise soon and instead linking to dictionaries seems to be the more feasible solution, and has the advantage that the creation of multilingual lexical data

[1] The top 10 languages by average ranking in GDP and number of speakers are: English, Chinese, Spanish, Japanese, Arabic, German, Portuguese, Hindi, Russian and French

[2] Such methods are required by *intelligent personal assistants* such as Apple's *Siri*

is now performed by those who have the interest and knowledge in doing so, instead of being a requirement on all data providers.

Based on the assumption that we need to link to entities defined in dictionaries and lexica, it is clear that there is some need for standardisation to define how this linking is performed and more importantly what format should be expected at when dereferencing such a link. This could happen in likely two ways: either the creation of a single large data source, likely based on a community based Wiki interface, causing a *de facto* standardisation, or preferably a standard format, introduced by organisations such as W3C or ISO, that allows for a competitive but inter-operable ecosystem for the description of such multilingual data. As such we[2] have proposed such a model we call *lemon*, the “Lexicon Model for Ontologies,” that aims to allows ontologies and linked data in existing semantic formats such as OWL and RDFS to be linked to rich lexical descriptions. We have continued to develop this as part of the OntoLex community group[3], with the aim of creating a linguistically sound model that will provide a guiding paradigm for producers of linked data lexica.

References

[1] Basil Ell, Denny Vrandecic and Elena Simperl (2011). Labels in the Web of Data. In Proceedings of the 10th International Semantic Web Conference.

[2] John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, Tobias Wunner (2012). Interchanging lexical resources on the Semantic Web. In: Language Resources and Evaluation, DOI:10.1007/s10579-012-9182-3.

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

Localization and interlinking of Semantic Web resources

Elena Montiel-Ponsoda and Guadalupe Aguado de Cea (Universidad Politécnica de Madrid)

Some of the most important challenges in providing multilingual access to the Semantic Web (SW) are related with two aspects:

1. The **provision of multilingualism** to ontologies and data sets documented in one natural language (NL)
2. The **interlinking** or **mapping** of semantic resources documented in different NLs

As the Open Linked Data phenomenon has shown, more and more resources are published in languages other than English [1]. A truly multilingual access to the SW involves, in our opinion, either the **localization** or translation of some resources to several NLs, or the **establishment of links** between and among ontologies and data sets in the same domain described in different NLs.

The main problem in the localization or interlinking matter is the fact that ontologies or data sets in the same domain may present some of these aspects:

- conceptualization mismatches
- different levels of granularity
- different perspectives of the same domain

Some of these aspects are also present in the interlinking of resources available in the same language, or what is the same, in the interlinking or mapping of **monolingual resources**.

In the **localization** of resources, current solutions fall short because of several reasons:

1. No homogeneous representation mechanisms accepted by the community are available. In this sense, several ontology-lexicon models proposed in the last years have tried to overcome this problem (LIR [2], *lemon* [3]).
2. Solutions fall short of accounting for conceptualization mismatches

We argue that in the localization of ontologies, specific representation models have to be able to define specific *relations* between NL descriptions in different languages, what we term *translation relations* or *cross-lingual relations*.

Highly related with this issue is the representation of **term variation** at a monolingual or multilingual level. A term variant has been defined as "an utterance which is semantically and conceptually related to an original term" [4]. We believe that the representation of term variants would also contribute to the establishment of links or relations between the NL descriptions associated to concepts (within or across languages).

A further problem may involve the automation of the localization process (See some proposed approaches for the automatic localization of ontologies [5] and [6]).

As for the **interlinking** of resources in different NLs, there are no specific links or mappings that can account for conceptualization mismatches between or among resources in several NLs. Some available solutions could be:

1. The “equivalence” or “sameAs” link would represent a solution in the case that highly similar conceptualizations are available for the same domain in different languages.

2. The “skos:broader” or “skos:narrower” link would work in some cases but their semantics are not clearly defined.

In both cases, the localization and the interlinking, in order to reach a principled solution we would need to provide well defined representation mechanisms and mappings intended principally to account for the differences between conceptualizations in different NLs. Without any doubt, standardization can play a key role to help solving this matter.

[1] Montiel-Ponsoda, E., Gracia, J., Aguado de Cea, G. & Gómez-Pérez, A. Representing Translations on the Semantic Web. *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW 2011),CEUR Workshop Proceedings, 775* (2011).

[2] Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., Peters, W.: Enriching Ontologies with Multilingual Information. *Journal of Natural Language Engineering*, 17 (3), 283{309 (2010).

[3] McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging Lexical Resources in the Semantic Web. *Language Resources and Evaluation*, in press (2011).

[4] Daille, B., Habert, B., Jacquenim, C., and Royauté, J.: Empirical observation of term variations and principles for their description. Terminology 3(2):197-257.

[5] Espinoza, M., G_omez-P_erez, A., Mena, E.: Enriching an Ontology with Multilingual Information. In *Proceedings of the 5th Annual of the European Semantic Web Conference (ESWC08)*, pp. 333-347 (2008).

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

Multilingual Word Sense Disambiguation and the Semantic Web

Abstract – Dagstuhl Seminar on the Multilingual Semantic Web

Roberto Navigli – Sapienza University of Rome

Motivation. The Web is by far the largest information archive available worldwide. Seen as a vast repository of text, the Web contains the most disparate information which can virtually satisfy all the possible user needs. However, nowadays the textual information needed by a user, such as in news, commentaries and encyclopedic contents, is provided in an increasing number of languages. For example, even though English is still the majority language, the Chinese and Spanish languages are moving fast to capture their juicy Web share, and more languages are about to join them in the run. The prototypical example of this trend is Wikipedia, whose multilingual growth is clearly exponential[1].

However, the Web suffers from two important issues:

1. First, the vast majority of textual content is not linked to existing ontologies, because of:

    i) the paucity of ontologies for several domains,

    ii) the lack of a suitable lexicalization for the concepts within many existing ontologies.

    While much effort has been devoted to the creation of ontologized information, the current state of the art is still very far from tackling the lack of domain and lexical coverage.

2. Second, the truly multilingual nature of today's Web is currently a barrier for most users, rather than an opportunity for having more and richer information. For instance, recently Google announced that Google Translate features about 200 million users per month[2], many of which are using mobile devices to obtain the appropriate lexicalization of their information need in another language. This need is also testified by the yearly organization of cross-lingual Information Retrieval forums like CLEF[3].

These key issues cry for the need of frameworks, tools and algorithms aimed at addressing the interactions between ontological representations and a babel of languages, so as to provide smooth access to multilingual content on the Web. The beneficiaries of such seamless integration would not only be end users, but also SMEs in virtually all industry sectors whose business is connected to the Web. In fact, an infrastructure able to overcome the language barrier would open new business opportunities in any domain, by increasing the customer base and approaching markets in new countries and regions.

Today's research in brief. The two main research communities concerned with the above-mentioned issues are, on the Web side, the Semantic Web (SW) community, and, on the language side, the Computational Linguistics (CL) community. On the one hand, the SW community has conducted much work in the direction of addressing the tasks of ontology construction, learning and population [2, 1], ontology linking [10], ontology matching [13], etc. On the other hand, the CL community has increasingly been working on important issues such as multilinguality [12, 4, 11, 8], disambiguation [6] and machine translation [5].

[1] http://meta.wikimedia.org/wiki/Wikimedia_in_figures_-_Wikipedia

[2] http://mashable.com/2012/04/26/google-translate-users/

[3] http://www.clef-initiative.eu/

What comes next. As suggested above there is an important missing link between the two communities, i.e., integrating ontologies with languages. Important efforts in this direction are DBPedia, YAGO, WikiNet, MENTA and Freebase. However, none of these proposals aims at bringing together the two worlds of the SW and CL by jointly and synergistically addressing the issues of ontological solidity, multilinguality and sense ambiguity. For instance, DBPedia is mainly concerned with popular types of Named Entities and manually maps concepts to WordNet, YAGO maps Wikipedia entities to the first senses of WordNet lemmas, MENTA addresses the multilinguality issue, focusing on the taxonomical aspect of an ontology, etc. In my group, an ambitious project – funded by the European Research Council – is currently under way, with two main goals: first, creating BabelNet [8, 9], a large, wide-coverage multilingual semantic network for dozens of languages and with many kinds of semantic relations; second, using BabelNet to semantically annotate free text written in any of the covered languages, thus performing multilingual Word Sense Disambiguation in arbitrary languages. This second goal is not addressed in other projects, and represents a step towards the multilingual Semantic Web. Still, the connection to the SW world is weak. In my vision, the next step is to **link data according to ontologies which provide multilingual lexicalizations**, a direction which I believe should be vigorously pursued in the near future and which would see the strong synergy of multilingual Word Sense Disambiguation with Linked Data. Here I do not mean we should create a single, global lexicalized ontology for all purposes. Instead, domain ontologies – even those which are not lexicalized according to a specific language – could be linked to multilingual lexicalized ontologies, which would be used as an interlingua for making Web content accessible to users independently of the language they master. Crucially, the more data will be linked across languages, the better the disambiguation performance (see e.g. [7]). One current problem of the Web of Data, in fact, is the ambiguity (and multilinguality) of labels for Linked Data [3]. W3C standards should be used to encode both the interlingual ontologies and the domain ontologies in a common format, and extensions of existing standards could be developed in order to bring together ontologies and the lexical meanings expressed in a babel of languages.

References

[1] Paul Buitelaar and Philipp Cimiano, editors. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, 2008.

[2] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors. *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123. IOS Press, The Netherlands, 2005.

[3] Basil Ell, Denny Vrandecic, and Elena Paslaru Bontas Simperl. Labels in the web of data. In *Proc. of International Semantic Web Conference 2011*, pages 162–176, 2011.

[4] Alexandre Klementiev and Dan Roth. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the ACL-COLING 2006*, pages 817–824, Sydney, Australia, 2006. Association for Computational Linguistics.

[5] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 2010.

[6] Roberto Navigli. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[7] Roberto Navigli and Mirella Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692, 2010.

[8] Roberto Navigli and Simone Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL-10*, pages 216–225, 2010.

[9] Roberto Navigli and Simone Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, to appear, 2012.

[10] Rahul Parundekar, Craig A. Knoblock, and José Luis Ambite. Linking and building ontologies of linked data. In *Proc. of International Semantic Web Conference 2010*, pages 598–614, 2010.

[11] Alexander E. Richman and Patrick Schone. Mining Wiki Resources for Multilingual Named Entity Recognition. pages 1–9, 2008.

[12] Yutaka Sasaki. Question answering as question-biased term extraction: A new approach toward multilingual QA. In *Proceedings of ACL 2005*, pages 215–222, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[13] Dennis Spohr, Laura Hollink, and Philipp Cimiano. A machine learning approach to multilingual and cross-lingual ontology matching. In *Proc. of ISWC 2011*, pages 665–680, 2011.

A very brief position statement on the multilingual web

Sergei Nirenburg

1) What are in your view the most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web?
There are many facets to this issue – technological, sociological, bureaucratic, etc. I can comment only on what the field of CL/NLP can contribute to the quality of the multilingual content. From this point of view the main challenge is reaching the level of quality of automatic translation that is acceptable by users. Manual translation is (currently) too slow to support fast turnaround on the Web. The current quality of automatic translation is too low for publication-level texts, though the general opinion is rather more favorable with respect to translation of informal texts, such as blog entries or tweets.
The above reckoning is not new and held for the pre-web machine translation applications as well.

3) Why do current solutions fail short?
See above: the low quality of the translated content. All other shortcomings are easier to overcome but in the final analysis they are not true obstacles.

4) What insights do we need in order to reach a principled solution? What could a principled solution look like?
In the short term, it is probably best to study the minimum levels of quality that users accept in various types of uses of the multilingual web and try to go after the low-hanging fruit first (not that much else has been going on in NLP over the last 15 years or so).
A gradual way of enhancing translation quality is – again, just like in the times of pre-web MT – human-assisted translation. It is not clear that this option would work for the web – after all, response speed is paramount in this environment. But if a successful methodology can be developed and if it can be shown to keep the costs of human translation down, then human-assisted translation can be a very useful stop-gap solution while more R&D is undertaken to develop high-quality automatic translation. Of course, there is no guarantee that this development period won't take several lifetimes.

5) How can standardization (e.g. by the W3C) contribute?
I don't think this is a crucial issue either way. In general, I think that standards should evolve and not be propagated top-down by bureaucratic means.

DAGSTUHL SEMINAR ON THE MULTILINGUAL SEMANTIC WEB

Under-resourced Languages and the Multilingual Semantic Web

Laurette Pretorius, University of South Africa

Challenge 1: Under-resourced Languages

In the context of the Multilingual Semantic Web (MSW) the most *basic* challenges that face Africa, and in particular South Africa, is limited internet access (out of scope here) and the proliferation of mainly under-resourced languages (in terms of financial, linguistic and human resources) with rich cultural diversity and indigenous knowledge systems encoded in these languages. Multilingualism impacts all sectors of African society, viz. public, private, business, education, etc.

As a first step, the Semantic Web (SW) may serve as a safe repository for valuable, already available language data/material by publishing it in the SW. To enable this process of preservation and archiving, the required range of language specific approaches, tools and technologies have to be developed. Care should also be taken to use or adapt, where possible, existing approaches and solutions in order to fast-track these initiatives. Time is of the essence in ensuring that these languages are technologically enabled.

In parallel to these archiving and repository building initiatives, the greater promise and benefits that the SW may offer to Africa, as it moves towards participating in the 21st century knowledge economy, should be immediately pursued. This would require the development of new terminology, state of the art lexical and language processing resources, tools, and technologies for the relevant languages. In addition, a wide and growing range of semantic technologies and ontologies may be required to capture the cultural diversity, the plethora of indigenous knowledge systems and all that goes with moving towards global economic participation and growth.

Challenge 2: Notions for clarification

The notions of “language-independent”, “culture-independent”, “language-specific”, “culture-specific”, the conceptual versus the linguistic, and how information about all of these is represented in the MSW, require continued careful consideration – a problem of which the complexity increases with the number and diversity of languages and cultures included.

Challenge 3: Interoperability and ease of use

The representation of the above notions, the information they are employed to encode, and, eventually, the resulting computational semantic artefacts (e.g. localised, mapped, modularised, verbalised ontology, etc.) will have to interoperate and they will have to be accessible across languages and cultures at a grand scale. At a more modest scale, for the real up-take of emerging semantic technologies and the MSW, it should also be relatively easy for a single user as producer and consumer of specialised content to conceptualise his/her arbitrarily complex interest domains, tasks, and applications, and to use the range of available MSW resources, representation and reasoning tools, etc. to his/her competitive advantage.

Examples of specific functionalities that may be relevant for a wide range of MSW users include:
(i) to have access to state of the art support and best practices of knowledge representation;
(ii) to do sophisticated intelligent searches of specified scope;
(iii) to delimit the search, access, generation and publication of information in languages of choice;
(iv) to perform automated reasoning of specified scope and complexity in the MSW;
(v) to obtain semantically accurate translations of the retrieved or generated material and of the reasoning results, on request;
(vi). to provide large-scale automated decision-making support in (multiple) natural language(s);
(vii). to have access to approaches and tools to evaluate results obtained.

Challenge 4: Continued interplay between natural language processing resources and technologies, semantic technologies and the MSW:

A serious issue in under-resourced languages remains the lack of terminology. The MSW offers unique opportunities in terms of community-based (crowd-sourcing) approaches to, among others, terminology development and moderation, and representations of culture-specific and indigenous knowledge systems. The MSW may serve as an incubator for the continued development of increasingly sophisticated natural language processing and lexical resources for under-resourced languages using new approaches that may emerge due to the availability of rich cross-language support, resources, tools and technologies.

Conclusion: The balance between appropriate theory and practice will be important in ensuring the sustainability of the MSW. Standards already exist or are in the pipeline for various aspects of the MSW. The ever increasing size and complexity of the MSW will require good standards and best practices towards some notion of integrity of the MSW.

Multilingual Semantic Web: Much More is Wanted

Aarne Ranta
www.cse.chalmers.se/∼aarne

Dagstuhl, September 2012

The crucial idea of the Semantic Web is the formalization of information. The advantage of formalization is that search and reasoning are supported by a well-defined structure: it enables us to work with formulas rather than text strings. This kind of information is easier for machines to process than free text. But it also involves an abstraction from languages and can thereby serve as a representation that supports multilinguality.

From the perspective of GF (Ranta 2011), the a formal structure can be seen as an abstract syntax—as an interlingua, which represents the content in a language-neutral way and makes it possible to produce accurate translations. This perspective is currently exploited in the European MOLTO project (Multilingual On-Line Translation).

The most important challenge for the Multilingual Semantic Web is simply: how to create more of it. The formalization is a wonderful thing when it exists, but it is expensive to create. We the supporters of the idea should try to make it more attractive—both by demonstrating its usefulness and by providing tools that make it easier to formalize web content.

MOLTO's emphasis has been on showing that formalization can give us high-quality automatic translation. MOLTO's show cases have been a small set of domains, ranging from mathematical teaching material to a tourist phrasebook. We have created techniques for converting semantic formalizations to translation systems, which cover up to 25 simultaneous languages. The richest demonstration of the idea is the Multilingual Semantic Wiki, built on top of the ACE Wiki (Attempto Controlled English) by generalizing it to a multilingual Wiki using GF. In this Wiki, every user can edit a document in her own language, and the changes are immediately propagated to all other languages. The translations are thus kept in synchrony at all times. This is enabled by the "master document" which is a formal representation of the content.

The main remaining challenge is: how to make new domains, new languages, and new information accessible in the Multilingual Semantic Web? We need to make it much easier to formalize legacy information, so that we can increase the amount of web content that can be accessed in the rigorous and reliable way. This formalization can benefit from heuristic bootstrapping methods such as machine learning—but it will always involve an element of human judgement, to make sure that the results are correct (Kay 1997). The challenge is to find the proper place of human judgement, so that its need can be minimized. The goal is to do more formalization with less effort.

The questionnaire

1. *What are in your view the most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web?*
Extending the coverage of the Semantic Web and the associated language resources.

2. *Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?*
High-quality translation: software localization, e-commerce, technical content, legal information, etc. Also language-based interaction and information retrieval, including mobile speech applications.

3. *Which figures are suited to quantify the magnitude or severity of the problem?*
The number of "concepts" involved (1000's in the largest cases of MOLTO; millions in the whole web); the number of languages (up to 25 in MOLTO, thousands in the world).

4. *Why do current solutions fail short?*
Too much and too boring human work is needed; its usefulness has not been convincingly demonstrated.

5. *What insights do we need in order to reach a principled solution? What could a principled solution look like?*
We need to understand what is easy (automatic) and what is difficult (needs human labour). We need to create logical and linguistic resources of high quality, coverage, and reusability, with completely free access.

6. *How can standardization (e.g. by the W3C) contribute?*
By eliminating duplicated work. For instance, if there is an ontology and multilingual lexicon of fish names, everyone can use it off the shelf and don't need to build their own from scratch. Having a common format is less important. It is OK to have different formats as long as they are fully specified and can be converted to each other.

References

ACE Wiki, http://attempto.ifi.uzh.ch/acewiki/.

GF, Grammatical Framework, http://www.grammaticalframework.org/.

MOLTO, Multilingual On-Line Translation, http://www.molto-project.eu/.

M. Kay, The Proper Place of Men and Machines in Language Translation, Machine Translation 12, pp. 3-23, 1997.

A. Ranta, Grammatical Framework: Programming with Multilingual Grammars, CSLI Publications, Stanford, 2011.

Author: **Kimmo Rossi**, European Commission, DG CONNECT, Unit G3 "Data value chain" (funding agency for research and innovation projects in the area of information and data management and language technologies).

This seminar comes at an interesting point in time, when we at DG CONNECT are defining the European Data value chain strategy, in view of establishing orientations for the first work programmes of Horizon 2020 (H2020), which is the next framework funding programme supporting research and innovation in ICT. Very soon after the seminar, we will launch consultations of stakeholders (researchers, industry, civil society, administrations) that will feed into the process. In early 2013 we need to have first stable topical orientations for the first phase of Horizon 2020. I expect this seminar to detect, define and refine some of the key methodological and scientific issues and challenges related to the data challenge in general, and the linked data/semantic web/text analytics challenge in particular. If it does, it will provide valuable input to the process of defining H2020 and other related programmes. With the recent reorganisation by which DG INFSO was rebaptized DG CONNECT, th e units responsible for information management (E2) and language technologies (E1) were combined into one unit. This created a single pool of over 100 ongoing research and innovation projects with over 300 MEUR funding, mobilising more than 1500 full time equivalents of Europe's best brains to break the hard problems that currently hamper effective use and re-use of online content, media and data. We try to use this pool of ongoing R&I projects also as a tool to bridge into the future research agenda in H2020, still taking shape. So, there are plenty of resources, the challenge is to make them converge and contribute to a common effort.

I also hope this seminar to be an opportunity of deepening the views and ideas that were presented at the Dublin workshop of the MultilingualWeb project, especially the 1st day, dedicated to the theme "Linked open data"
http://www.multilingualweb.eu/en/documents/dublin-workshop/dublin-program

Below are my personal views concerning the specific questions:

1) What are in your view the most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web? In general, the concept of Semantic Web (accessing and addressing the web as a database) requires precise, fast and robust automated text analysis capabilities which are not there, especially for languages other than English. Since the increasing majority (75% or so) of Web content is in languages other than English, the text analysis bottleneck gets worse over time.

2) Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem? Huge potential benefits are currently missed because of lack of semantic tagging and linking of documents. Such benefits could be reaped in various sectors, but the most obvious are publication, communication, marketing, intelligence, tourism, biomedical, administration -- any industry or activity which relies on fast access to relevant facts from large numbers of textual (human language) sources, some of which are static documents, other streams of data.

3) Why do current solutions fail short? I may be wrong, but I have a feeling that state of the art information extraction has not been systematically utilized in efforts to link data, for example in efforts like DBPedia. Also, there is too much hand-crafting, domain-adaptation, and other tweaking that is not replicable and robust. When the characteristics of the underlying data change, such schemes risk becoming obsolete. Another thing is that the performance of automated tagging, information extraction, named entity recognition etc. are heavily dependent on the language (and completely missing for some languages). Finally, methods and solutions should be able to cope with

Abstract for the Dagstuhl Seminar on Multilingual Semantic Web
Author: Felix Sasaki (DFKI / W3C Fellow)
Title: Sustainable, organisational Support for bridging
Industry and the Multilingual Semantic Web
Version: 14 July 2012

1) Challenges

A huge challenge for the multilingual Semantic Web is its separation to other types of multilingual content. In industries like localization, terminology or many areas of linguistic research, creation of multilingual resources is resulting in fast amounts of content in many languages. Unfortunately, this content cannot be easily re-used on the multilingual Semantic Web, and resources created within the multilingual Semantic Web are rarely part of industry applications. The underlying issue is partially one of information integration; this problem is already tackeled via formats like NIF[1] and LEMON[2], which help to re-use and combine tools and resources (e.g. lexica) for the creation of multilingual resources. However, another part of the issue is the topic of content creation and localization workflows, which is different in industries compared to the multilingual Semantic Web. This difference can be characterized as follows:

- In localization and content creation, multilingual resources are being created in complex workflows with many organizations involved. In the multilingual Semantic Web, multilingual linked open data is rather created on a project specific basis by research groups. This leads to uncertainty with regards to the quality and maintanance of the data.
- Thurstworthiness and quality of data is an important aspect of workflows in localization and industrial content creation: e.g. the localization of medical information needs to take national and – esp. in translation scenarios – international regulations and quality measures into account. Data currently available on the multilingual Semantic Web not only differs highly in terms of quality; an evaluation of the quality itself (level of maintenance, trust of content creators etc.) is hard to achieve.
- Closely related to thurstworthiness are legal aspects of linked data, e.g. what data can be re-used with what kind of license. Without such information, data from the linked open data cloud will not be re-used in large scale industrial scenarios.

2) The role of the problems in industry practice

The above problems matter in practice since so far the usage of linked open data in areas which are inherently multilingual, that is content creation and localization, is rather low. This does not only have to do with current technical solutions for the multilingual Semantic Web itself: as Jose Emilio Labra Gayo (2012)[3] demonstrates, in the current technical infrastructure there are already means to create multilingual information within linked open data; unfortunately these are rarely used, and the actual amount of multilingual data in the Semantic Web is rather low.

3) Why do current solutions fail?

Technical advances are a mandatory part of a solution to the problem, see e.g. LEMON and NIF mentioned above. Nevertheless, failures are also due to organizational issues.

An example of this situation are language identifiers and what I will call the "zh" problem. "zh" is the language identifier for Chinese created as part of ISO639-1. The first edition of ISO639-1 was approved 1967; here "zh" is described as an identifier for Chinese in general. However, for decades, it has mainly been used for identifying content

in the Mandarin language. With the creation of ISO639-3, Mandarin received its own language identifier “cmn”. “zh” was defined as a so-called macrolanguage, acknowleding that there is no single Chinese language. “zh” now is a macrolanguage covering closely related languages like Mandarin or Hakka.
The new role of “zh” leads to the situation that a language tag in existing content like “zh-tw” can have several interpretations: the macrolanguage Chinese spoken in Taiwan, Chinese in the traditional script, or Mandarin in Taiwan.

4) What insights do we need in order to reach a principled solution?
The lesson to be learned from “zh” is that what is needed are not only multilingual resources, e.g. the identifier “zh”, or advances in technical solutions. In addition, organisational and workflow information about the context of content creation and applications need to be established.
Various industries (libraries, terminologists, general software companies, Web centered companies and multimedia companies) are working closely together to solve problems which arise from the above situation, that is: to make sure that in a given workflow, “zh” can be interpreted in the appropriate manner.
Currently the community of multilingual Semantic Web is not part of the related organizational structures, which creates barriers between the multilingual Semantic Web and other, inherently multilingual industries. The bad news is that there is no general, principled solution to resolve this. But we can make steps and long-term plannings which will help to address the problem.

5) How can standardization contribute?
Standardization is one important part to solve the workflow problems described in this abstract. The community building needed to solve the “zh” problem for the industries mentioned above mainly is happening in standardization bodies like the IETF, the Unicode consortium or W3C. The aforehand mentioned efforts of LEMON and NIF show that the NLP community is making efforts into the direction of standardization. The W3C MultilingualWeb-LT Working Group[4] is another effort with special focus on community building involving industries, making sure that there is awareness for issues mentioned under 4).
Nevertheless, the conclusion is that this is not enough: what is needed is an approach also towards research in which integration with industry workflows is not an aftermath or part of separate projects. Sustainable, institutional support for bridging the workflow related gaps mentioned in this abstract is needed. We should put an effort in describing research and (product) development not as two separate lines of action, but as closely integrated efforts. How this sustainable instutionalization of multilingual research and innovation should be framed in detail and how it should be worded in upcoming research programs like Horizon 2020, is an important and urgent topic. The Dagstuhl seminar should help to move this discussion forward, also just by bringing the relevant players together.

[1] NIF (NLP Interchange Format) http://nlp2rdf.org/

[2] LEMON (LExicon Model for ONtologies), being standardized in the W3C Ontology-Lexicon community http://www.w3.org/community/ontolex/

[3] Best Practices for Multilingual Linked Open Data. Presentation at the 2012 workshop “The Multilingual Web – Linked Open Data and MLW-LT Requirements”, Dublin. See http://www.multilingualweb.eu/documents/dublin-workshop/dublin-program

[4] See http://www.w3.org/International/multilingualweb/lt/ for further information.

Dagstuhl Seminar on the Multilingual Semantic Web

Short position statement

Gabriele Sauberer, TermNet

1) Most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web:

The main problem of the Semantic Web and the Multilingual Semantic Web (MSW) alike is the imbalance of its players and drivers, i.e. the lack of diversity: they are mainly male, white, academic, IT-focussed and aged between 20 and 45.

It´s not only a matter of dominance of "English language and Western culture" as correctly stated in the Synopsis of the Dagstuhl Seminar, it is much more a matter of a global digital divide: a divide between men and women, between age groups, social classes, between disciplines, subject fields, traditions, cultures, information and knowledge rich and information and knowledge poor, between literates and illiterates, experts and non-experts, etc.

> *Thus, one of the pressing needs with respect to the MSW is to address the lack of diversity and to overcome the global digital divide.*

2) Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?

Lack of diversity in working together to build a Semantic Web that matter for all citizens is, to my mind, one of the main reasons why MSW got stuck – technically, economically and socially.

The word-wide acceptance of the Semantic Web is a key issue of its development and survival. People all over the world understood the benefit and practical advantages of mobile phones in their lives very fast.

> *What´s in it for all of us when using semantic web technologies and smart phones is the core message to be brought home by the drivers of the MSW.*

There are many industries which can help in overcoming language and national barriers: the language industries, education and training industries, Information and Communication Industries, etc. "Facilitating semantic access to information originally produced for a different culture and language" as mentioned in the synopsis is to be avoided, to my mind, not fostered. Why? Because goal and vision of MSW should be to empower people to contribute to the MSW by their own in their own languages, not being restricted to adept and localize foreign content.

> *Terminological methods, tools, trainings and consultancy services are, to my mind, key technologies and basic knowledge prerequisites to contribute to problem solutions.*

3) Which figures are suited to quantify the magnitude or severity of the problem?

The industrial relevance of MSW and its barriers is high-lighted at page 3 of the synopsis, e.g.:

> Especially in knowledge-intensive domains, such as finance, biotechnology, government and administration etc., the ability to interface with Semantic Web or Linked Data based knowledge repositories in multiple languages will become of increasing importance. In finance, knowledge repositories will be build up of company-related information, i.e. in terms of finance, markets, products, staff, all of which will be curated and accessed in multiple languages.

> *No doubt, we are talking here about a hundreds of Billions Dollar, Euro, RMB etc. business / losses.*

4) Why do current solutions fail short?

> *Because the current solutions are no sustainable, future-oriented and no global solutions: They lack of creativity and innovation, caused by lack of diversity. It´s just more of the same, provided by the same players (see question 1).*

5) What insights do we need in order to reach a principled solution? What could a principled solution look like?

What we need is, to my mind, exchange of insights at all levels, representing the diversity of current and future users of MSW (all genders, all age and social groups, all regions, cultures, disciplines, literates, illiterates, etc.).

To seize opportunities for new insights is crucial and simple, but not easy: We mainly need to overcome the barriers and restrictions in our minds, in our ethno-centric attitudes and behaviour.

> *A principled solution could be to make diversity a main principle of the Semantic Web and the MSW: With this new and lived principle, diverse teams and diverse expert and working groups, guided by communication and terminology experts could make the vision of a real multilingual and multicultural Semantic Web come true.*

6) How can standardization (e.g. by the W3C) contribute?

> *Standardization organizations and their Technical Committees as well as the W3C can contribute to the principled solution by developing, issuing and promoting respective standards and guidelines to organize and support diversity as main principle of the Semantic Web and the MSW.*

Hans Uszkoreit

The Translingual Web

A Challenge for Language and Knowledge Technologies

Abstract of a Position Paper for the Dagstuhl Seminar on the Multilingual Semantic Web

The web becomes more multilingual every day and so does the society of web users. This is not surprising since many large organisations and societies are already multilingual in their processes and constituencies. Multinational corporations, international organisations, professional associations as well as national and regional societies such as the European Union, South Africa, India and Russia often work in many languages.

The World Wide Web has become the predominant medium for sharing information, knowledge and artistic content. The Web is also turning into the universal platform for an endless number of services that extend the static content of the web by functionalities using or modifying this content or just utilizing the Web protocols for all kinds of transactions.

There is a strong demand for making the fast-growing multilingual web also truly crosslingual so that the global medium, which unites all the contents and services in more than thousand languages, would also eventually become the instrument for overcoming all language barriers. So-called globalized websites and web services today offer contents and services in 20-35 languages. Such websites are hard to maintain, especially if the contents grow and if the services depend on reliable translation. Google translate offers translations from and into 57 languages. The popular translation facility is an invaluable service making Internet content accessible to large parts of the world's population that would otherwise be deprived of the Web's blessings. However, well-known quality deficiencies of today's translation technology limit the role of the existing large online translation services as the universal crosslingual information and communication hub.

The semantic web is an ambitious program driven by a powerful vision and a promising approach toward a web of knowledge-based services that become much less dependent on human language and therefore also on human languages. If the entire Web could be encoded in semantic representations that are language-independent and suited for automatic reasoning, the crosslingual function of the Web would boil down to multilingual access facilities. The main challenge for such a natural language interface would be the understanding of spoken or written queries and commands. Compared to this unsolved central problem of language processing, the non-trivial task of responding in any requested language is comparatively simple, as long as the semantic web services select the appropriate output in a structured representation.

However, in the foreseeable future we will not witness a Web in which all content is represented in a disambiguated structured semantic representation. At best, a growing layer of semantic web content and services will sit above the wealth of unstructured content, containing large numbers of links into the texts (and possibly also other media) that let the extracted knowledge also serve as metadata for the unstructured information base. Two observations: (i) the evolution of this semantic layer proceeds in ways not quite foreseen by the early visionaries of the Semantic Web and (ii) the resulting large heterogeneous distributed metadata repository may soon become the most important research and technology resource for getting at the meaning of so-called unstructured data, especially texts for which such metadata do not yet exist.

After more than 50 years of research with human language processing, the scientific community has learned from a combination of experience and insight that in this discipline there are no miracles and no free lunches. Neither teraword data nor a century of grammar writing, neither fully automatic learning nor intellectual discovery and engineering alone will suffice to get us the technology that reliably transforms language into meaning, or one language into another language. It will not even give us the tools for correctly and exhaustively transforming every meaningful linguistic utterance into its paraphrases in the same natural language. Even treebanks including parallel treebanks for several languages with their sophisticated structural information will not provide a sufficient empirical basis for the maturation of language technology. The recognition of the need for semantically annotated textual data keeps growing. Even large semantic resources such as Yago or Freebase whose knowledge units are not directly linked to the texts they came from, have proven highly valuable in language technology research, especially in information extraction.

But the need for semantic resources also includes translation technology. After hierarchical phrase-based and syntax-based translation, semantics-based statistical translation will become the next big trend in MT. On the other hand, knowledge technologies will not be able to get into full blossom either without the evolutionary upgrade path from the textual knowledge representation to the semantic metadata layer.

It seems that the prospects of language technologies and knowledge technologies are inseparably tied to together. Since each of the two technology disciplines needs the other one for reaching fruition, only a complex mutual bootstrapping process around a shared stock of data, tools and tasks can provide the base for effective evolution. A Multilingual Semantic Web layer could become the shared resource of data and tasks for this process, if the Multilingual Semantic Web indeed becomes a core component of the envisaged Translingual Web, it could also incorporate the services needed on the NLP side. Such services would not only cater to research but also gradually fill the growing cross-lingual needs of the global multilingual society.

Both in language technologies and in knowledge technologies research has become much more interconnected and collective in nature. As in the natural sciences and other engineering disciplines, new forms of collaboration and sharing have developed. The sketched bootstrapping will require additional efforts in sharing challenges and resources. How could such efforts be triggered and organized? For European language technology community, some important planning steps toward large-scale cooperation have been taken.

Coordinated by the Multilingual Europe Technology Alliance (META), the European language technology community together with technology users and other stakeholders has drafted a Strategic Research Agenda (SRA), in which the special needs and opportunities for language technology in our multilingual European society are outlined that should drive our research. From these findings, the SRA derives priorities for large-scale research and development as well as a plan for implementing the required massive collaboration. Among the priorities are three solution-driven research themes that share many technologies and resources. All three priority themes are tightly connected with the topic of the Dagstuhl Seminar: (i) a cloud computing centered thrust toward pervasive translation and interpretation including content and service access in any language, (ii) language technology for social intelligence supporting participatory massively collective decision processes across social, linguistic and geographic boundaries and (iii) socially aware proactive and interactive virtual characters that assist, learn, adapt and teach.

As a means for experimentation, show-casing, data-collection, service integration and actual service provision, a sky-computing based platform for language and knowledge services is planned that needs to be realized through a cooperation between industry, research and public administration. This platform would be the natural target for experimental multilingual semantic web services.

The interoperability of the services will to a large degree depend on the success of ongoing standardization efforts as conducted in collaborations among research, industry, W3C and other stakeholders in the framework of the initiatives *Multingual Web* and its successor *LT-Web.*

Besides the valuable exchange of recent research approaches and results, the Dagstuhl seminar could play an important part in a better linking of the following five research areas in planning:

1. Semantic Web Standards and methods
2. Linked open data and other knowledge repositories
3. Multilingual Web Standards
4. Translingual (Web-based) Services
5. Information/Knowledge Extraction

In addition to new developments from the META-NET/META community (visions, strategic research agenda, schemes for distributed resource sharing) I will try to contribute experience and perspectives to this endeavour from two specialized research strands: (i) minimally-supervised and distantly supervised n-ary relation extraction and (ii) quality-centered translation technology.

Problems and Challenges Related to the Multilingual Access of Information in the Context of the (Semantic) Web

Josef van Genabith, Centre for Next Generation Localisation and National Centre for Language Technologies, Dublin City University

The topic of the Dagstuhl Seminar is broad, especially as the “Semantic” part in the title is in brackets, which could suggest optionality, as in “Web” or “Semantic Web”. Accordingly, the notes below quite broad (and rambling).

Challenges:

Challenges include (and go well beyond) a mixed bag of related philosophical (knowledge representation, epistemological, reasoning), granularity, coverage, multi-linguality and interoperability challenges.

Philosophical: the semantic web aims at capturing knowledge, mostly in terms of concepts and relations between concepts, to support automatic access to and reasoning over knowledge. However, the base categories are not clear. What is a concept? Do concepts exist independently of culture, language, time? Are concepts conventionalised, political, or even ideological constructs? Is it a matter of degree, is there a spectrum with extreme ends with pure concepts on the one hand and completely culturalised concepts on the other. If so, how do we know what is where on the spectrum and how does this impact on computation? Are ontologies in fact folksonomies etc.? Do we need to bother? Yes, as multi-linguality (amongst others) shows that concepts are not as universal as perhaps assumed. Reasoning (beyond relatively simple applications) is extremely challenging both computationally and conceptually: reasoning with events, temporal information, hypothetical information, contradicting information, factivity, sentiment, probabilistic information, causality, etc. Maybe ontological information (Semantic Web) should be extremely confined/demarcated (factual, as accepted by a culture) backbone component feeding into to a more general inferencing process (link with NLP).

Granularity: for many applications a shallow ontology is quite useful. There is a question when to use/compile shallow or deep/detailed ontological information and for which purpose (fit-for-purpose)? What is not covered by ontological information?

Evaluation: how do we evaluate ontological information? Is there an abstract measure, or is it just in terms of some usefulness criterion given a task (extrinsic evaluation)?

Coverage, quality, noise: content (un-, semi and structured) is exploding on the Web with ever increasing volume, velocity and variety. How do we obtain ontological information: manually compiled, automatically compiled from (semi-) structured input (tables etc.), or from raw text (through NLP/IE)? We need to negotiate the engineering triangle: cheap, fast, quality (you can only have two out of the three at any one given time).

Multi-linguality: most of the Semantic Web is in English. Multi-linguality raises issues including culture specificity, mapping between ontological information resources (which is quite alien given the often un-articulated background assumption that ontologies are about concepts that may help transcend languages and cultures), overall structure for ontological information (one concept for each culture and "translations" between them, a single one with sub-categorizations: Chinese, Arabic, Western ….)?

Interoperability: multi-lingual, -cultural, -granular, -redundant, -domain, -... how do we make this all play in concert? How do we make NLP/IE and Semantic Web interoperable? They should be able to contribute much to each other, in fact (some of) the trouble starts when you make each one of them do it all on its own ….

Why does problem matter in practice?

We need to capture knowledge to support technology mediated access to and interaction with knowledge.

Figures that quantify the problem:

Content velocity, volume and variability is steadily increasing: rise in User Generated Content (UGC) with Web 2.0 move of users from passive consumers to active producers of content. English is rapidly loosing its role/status as the language of the web. Most growth in the web is from emerging economies.

Why do current solutions fall short?

The trouble is there are different kinds of knowledge (of which the Semantic Web captures some), the volume of knowledge is constantly increasing (of which the Semantic Web captures some), knowledge is dynamic (i.e. constantly changing, updating) (of which the Semantic Web captures some), knowledge is encoded in different formats (un-, semi- and structured) (of which the Semantic Web captures some) and different languages (of which the Semantic Web captures some).

Principled solution:

In my view making Semantic Web and NLP play in concert supporting each other is one of the greatest challenges. NLP can provide scalability, the capacity to address content/information velocity (frequent updates), volume and variety. Semantic Web can provide knowledge guiding NLP. NLP can help bridge language barriers.

Standardisation:

Full standardization is difficult to achieve. It may be more realistic to aim for some kind of interoperability of heterogeneous sources of information/knowledge.

Be Informed and the Multilingual Semantic Web

Jeroen van Grondelle, Principal Architect, Be Informed
http://www.beinformed.com/, j.vangrondelle@beinformed.com

This contribution aims to provide an industry perspective on the multi-lingual semantic web and tries to answer three questions: What is the semantic web used for today, why is (natural) language important in the context of the semantic web and, only then, how could that guide the development of the multilingual semantic web.

A Semantic Web of Unconsolidated Business Constraints
For reasons that probably differ from the original semantic web vision, companies and governments alike are embracing semantic technology to capture the information they use in declarative, interoperable models and ontologies.

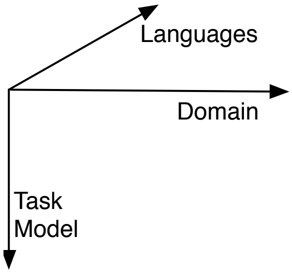
They move beyond the data and capture the policies and definitions that govern their daily operations. By choosing the right conceptualizations for business aspects such as products, business processes and registrations, they manage to use these ontologies to design their business, agree on the terms used to communicate its intentions and execute the required applicative services that support it.

The semantic web stack of ideas and technologies fits them well. Although the problem does not have global web scale, they benefit from the unconsolidated nature of the semantic web technologies when capturing aspects across different organizations and departments. The vocabularies that emerge when modeling this way have proven very valuable in communicating policy between all stakeholders.

A Lingual Semantic Web for Humans AND Machines
Although a lot of emphasis lies with machine's ability to interpret and reason with ontologies on the semantic web, we believe that human's role is crucial.

In our experience, ontologies can be useful to organizations **throughout the policy lifecycle**: From shaping the organization by drafting policy candidates and choosing the policy that is expected to meet the goals best, implement policy in business processes and applications, execute policy and evaluate policy by reporting and collecting feedback. When ontologies are used to facilitate these processes, users will interact with the ontologies in many ways: Domain experts and business users need be the owners of the ontologies and therefore need to create, review and validate the ontologies. and users need to interact with and understand the services based on the ontologies, ranging from online forms to complex process applications.

Language plays an import role in **supporting the different forms of dialog** needed. Often the boundaries of dialog supported by a specific language technology are specified at a lingual level, bound by lexicons and types of lexical constructs understood by the technologies etc.. When used in the context of the semantic web, we believe that the boundaries of dialog need to be specified in two dimensions: The domain that is discussed in dialog, typically represented by an ontology, and the **task context** of the dialog.



Often the implicit task model under semantic web applications is limited to querying and presenting instances in the ontology, or aggregates of them. Both the original semantic web vision and the businesses using semantic technology today require more complex tasks to be performed based on ontologies, and more diverse dialogs as a consequence.

For example, an ontology might be use to capture policies on eligibility for a permit [IND]. Typical task models throughout the policy lifecycle might require dialogs that

- Speak of the domain in general or that speak about a specific permit application;
- Allow citizens to apply for a permit, providing all required facts and receiving feedback on the consequences of the information they provide;
  - Discuss an application with an official or with the applicant himself;
- Support experts to validate the ontology and maintain consistency;
  - Represent not only the ontology, but also represent generated contradictions to trigger feedback.
- Support what if analyses, that describe possible changes in the legislation or in the population of applicants and the consequences on the permits issued;

All these types of dialog require ingredients both from the domain ontology and from the task model, probably at both a semantical level and the lexical representation level. Challenges in LT might include representing the different aspects of such a task model, and how to decouple it in an orthogonal way from the domain ontology.

Use Cases and Challenges for Multilingualism

Given the importance of language in interacting with the semantic web, it is clear that multilingualism is crucial when applying semantic web technologies at serious scale and in international context. Apart from providing transparent access to services and dialogs based on ontologies, multilingual capabilities of the semantic web are important for sharing and reusing ontologies and facilitate collaboration across languages in the process of creating and agreeing on ontologies that capture international standards.

The task orientation introduces requirements to multilingualism beyond translation of all concepts in both dimensions. There are a lot of lingual, even cultural aspects to having a dialog, such as when to use formal forms, what are the preferred ways to ask, confirm and give advice for instance.

Conclusion

Most language is spoken in dialog, in the context of a task or a common goal and in a certain domain. Language technology that incorporates the conceptualizations of all these aspects and is able to generalize across languages has useful applications today in lots of areas, including the business processes of both companies and governments. That may be a first step to making the semantic web vision a reality: a web of intelligent agents and services based on unconsolidated, distributed ontologies, created and owned by domain experts.

Position Statement for Dagstuhl Seminar on “Multilingual Semantic Web”

Martin Volk, University of Zurich July 13, 2012

The biggest challenge in multilingual access to the web is still the **limited quality of machine translation**. This may sound like a somewhat trivial observation, but it clearly points to the core of the problem. Machine translation has made big progress. Because of statistical machine translation we can build translation systems quickly for many language pairs when large amounts of translated texts are given for the languages and domains in question. The quality of machine translation in many application areas is good enough for profitable post-editing rather than translating from scratch. But the quality is often still a problem when using the machine output in other applications like cross-language information extraction.

Large collections of translated texts (parallel corpora) are the crucial prerequisite for advancing not only the field of machine translation but also any approach to automatically learn cross-language term correspondences and to verify and disambiguate ontological relations for each language. After all large text collections are the sole basis for automatically extracting semantic relations on a large scale.

Therefore I see it as our most important task to collect parallel corpora, to encourage more people to provide parallel corpora and to support any initiative for the free access to a wide variety of parallel corpora.

Still, we shall not forget that statistical approaches to translation and to ontology building are not an option for many **lesser-resourced languages** (which account for the vast majority of the languages spoken on the planet today). In order to work against the widening of the technological gap, we need to find viable approaches to build bilingual dictionaries with large and lesser-resourced languages, to collect special corpora for such language pairs and to include these lesser-resourced languages in our ontology building efforts.

All activities towards multilingual information access, in the web in general and in the semantic web in particular, will benefit **many types of industries** and organizations: the media industry (newspapers, TV stations, film industry), the manufacturing industry (user interfaces, user manuals, documentation), trade and commerce (communication, agreements, contracts), administration (law texts, court decisions, patents), tourism, science and education.