Lessons learned (and questions raised) from an interdisciplinary Machine Translation approach

Position paper for the W3C Workshop on the Open Data on the Web,
23 - 24 April 2013, Google Campus, Shoreditch, London

Timm Heuss
University of Plymouth, Plymouth, United Kingdom
Timm.Heuss@{plymouth.ac.uk,web.de}

March 2, 2013

Abstract: Linked Open Data (LOD) has ultimate benefits in various fields of computer science and especially the large area of Natural Language Processing (NLP) might be a very promising use case for it, as it widely relies on formalized knowledge. Previously, the author has published[1] a fast-forward combinatorial approach he called "Semantic Web based Machine Translation" (SWMT), which tried to solve a common problem in the NLP-subfield of Machine Translation (MT) with world knowledge that is, in form of LOD, inherent in the Web of Data. This paper first introduces this practical idea shortly and then summarizes the lessons learned and the questions raised through this approach and prototype, regarding the Semantic Web tool stack and design principles. Thereby, the author aimes at fostering further discussions with the international LOD community.

1 Motivation

There are arbitrary uses residing in the nature of every information and with 5-star-Linked Open Data (LOD) [2], humanity could finally let machines benefit from the knowledge that is available world wide, to everyone.

While all fields of computer science could ultimately benefit from the Web of Data, one of the most pressing area is probably the giant field of Natural Language Processing (NLP), which widely relies on formalized knowledge in various kinds, formats and applications. One example are dictionaries, which are usually designed to be in a very application specific or even proprietary format. In addition, world knowledge often plays an important role for the quality of a NLP system.

In the NLP-subfield of Machine Translation (MT), it is often crucial not only to have a comprehensive dictionary, but to understand the source text correctly, as otherwise ambiguities may result in incomprehensible target language translations [5].

Within the last decades, the field of MT has undergone various changes of underlying technology, from rule-based to statistical approaches. Especially the statistical approaches benefit from the vast amounts of information on the "eyeball Web" [3, p. 82].

However, none of these MT approaches reflect the knowledge-based developments the Web made in the realms of the Semantic Web respectively the Web of LOD since the millenium. In the opinion of the author, a logical next step for MT is to involve the knowledge that is inherent in the Web of Data. He therefore developed a combinatorial approach of the Semantic Web and MT disciplines, which he called Semantic Web based Machine Translation (SWMT) [5].

[1] `http://mt-archive.info/EACL-2012-Harriehausen.pdf` (accessed 2013-02-23).

2 Semantic Web based Machine Translation

The idea behind SWMT is basically to use the knowledge that is available in form of LOD as dictionary for a MT task. Thereby, a principle design goal is to use the vision and standards of the W3C as strict as possible: There are a few Resource Description Framework (RDF) statements containing the world knowledge, natural language is represented as `rdfs:label`, as a "human-readable version of a resource's name" [4], including a language notation following RFC-3066[2] [6]. In addition, reasoning by the Web Ontology Language (OWL) expands the knowledge logically. Thus, the approach works with implicit, inferenced knowledge that is not stated explicitly by RDF statements.

2.1 Sample scenario

To demonstrate the powerfulness of this approach, the following short but extremely tricky sentence was literally designed to be translated in the author's native language German:

> `Pages by Apple is a word processor like Word by MS.`

Of course, `Pages`, `Apple`, `Word` and `MS` are proper names and should not be translated. An additional measure to stress a machine translation is to use indirect product names (`Pages by Apple` and **not** `Apple Pages`). Thus, these phrases cannot be derived from possible dictionary entries. Also, the company `Microsoft` is abbreviated by the common `MS`.

For traditional translation approaches, sentences like this are usually extremly hard to translate [5, page 5]. But for SWMT, a translation is a logical conclusion of world knowledge.

2.2 Involved knowledge / LOD

The sentence mentioned above contains the world knowledge that a vendor called `Apple` produced a product named `Pages` and a vendor called `Microsoft` with the short form `MS` produced a product named `Word`.

Of course, DBpedia[3] is the first address to retrieve that knowledge from. For example, the following triples have been extracted to reflect world knowledge about `Microsoft`:

```
dbpedia:Microsoft_Word dbp:developer
    dbpedia:Microsoft .
dbpedia:Microsoft_Excel dbp:developer
    dbpedia:Microsoft .

dbpedia:Excel dbo:wikiPageDisambiguates
    dbpedia:Microsoft_Excel ; rdfs:label
    "Excel" .
<http://dbpedia.org/resource/Word_(
    disambiguation)> dbo:
    wikiPageDisambiguates dbpedia:
    Microsoft_Word ; rdfs:label "Word" .
dbpedia:MS dbo:wikiPageDisambiguates
    dbpedia:Microsoft ; rdfs:label "MS" .
```

The property `dbo:wikiPageDisambiguates` is refereed to add common short forms (e.g. `Excel` for `Microsoft Excel`) and abbreviations (e.g. `MS` for `Microsoft`).

This general-use LOD is supplemented with two application specific enrichments: First, the property `dbo:developer` gets an more expressive and especially internationalized human-readable label than DBpedia's `developer`:

```
dbp:developer rdfs:label "by"@en, "von"
    @de, "produziert von"@de .
:produces rdfs:label "produces"@en, "
    produziert"@de.
:produces owl:inverseOf dbp:developer .
```

[2] `http://www.ietf.org/rfc/rfc3066.txt` (accessed 2012-01-25).

[3] `http://dbpedia.org/About` (accessed 2013-03-01).

The second addition is the definition of application specific trigger words, which are used in the following section to intensionally trigger the enrichment of the translation dictionary out of LOD:

```
dbpedia:Microsoft_Word a :trigger .
dbpedia:Apple_Inc a :trigger .
```

2.3 Prototype

Since its publication last year [5], the prototype[4] has changed significantly. The basic principle, however, is still the same: The application translates an English sentence to German, by emulating a traditional translation system that translates words and phrases with the help of a (also very traditional) dictionary. The novelty of the approach is the Semantic Web tool stack, that uses the LOD to create new dictionary entries based on reasoning on world knowledge. The figure 1 depicts this combination of traditional and novel translation components.

As mentioned in the previous section, the DBpedia entries `dbpedia:Microsoft_Word` and `dbpedia:Apple_Inc` are defined as triggers. Thus, if a sentence contains words like `Word` or `Apple` (labels that are retrieved by referencing the `dbo:wikiPageDisambiguates`-relation), the Semantic Web tool stack is asked to produce additional dictionary entries for the translation. After creating an inferred model with an OWL-reasoner, this essentially involves a Simple Protocol And RDF Query Language (SPARQL) query. By the example of the trigger word `Pages`, this query looks like the following:

```
SELECT  ?sbLabel ?doesLabelFrom ?
    sthLabel ?doesLabelTo WHERE
  { ?sb ?does ?sth .
    ?does rdfs:label ?doesLabelFrom .
    ?does rdfs:label ?doesLabelTo .
    ?gSb dbo:wikiPageDisambiguates ?sb .
    ?gSb rdfs:label ?sbLabel .
    ?gSth dbo:wikiPageDisambiguates ?sth
        .
    ?gSth rdfs:label ?sthLabel
    FILTER langMatches(lang(?
        doesLabelFrom), "EN")
    FILTER langMatches(lang(?doesLabelTo
        ), "DE")
    FILTER ( regex(str(?sbLabel), "Pages
        ", "i") || regex(str(?sthLabel),
        "Pages", "i") ) }
```

[4] https://github.com/heussd/swmt (accessed 2013-02-28).

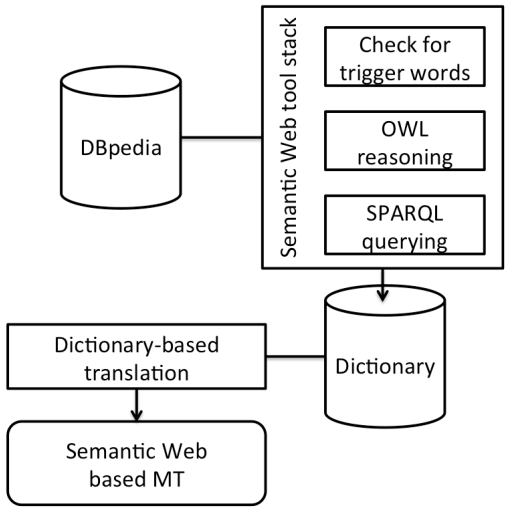

Figure 1: Prototype architecture consisting of a traditional, dictionary based translation mechanism and a corresponding dictionary. This dictionary is enriched with Semantic Web technology by reasoning and querying LOD from DBpedia.

As a result, simple but analogous and valid translations based on actual word knowledge are produced. In the given scenario, for the trigger word `Word` the phrases shown in the following figure are produced:

English	German
Word by MS	Word produziert von MS
Word by MS	Word von MS
MS produces Word	MS produziert Word

Figure 2: Valid translation phrases, reasoned and queried out of the LOD from DBpedia.

These phrases are then added to the dictionary. As a result, the translator produces a valid and analogous German translation for the given sample sentence that would have never been translated by traditional approaches:

```
Pages von Apple ist eine Text-
verarbeitung wie Word von MS.
```

This application is available at GitHub[5].

3 Lessons learned and questions raised

As mentioned, the prototype is designed as close as possible to the vision of LOD and the building paradigms of the Semantic Web. Connected to this design goal, some issues emerged during development. In the following, the author describes those issues and where he sees conceptual mismatches, either of the two domains NLP and LOD or in the visions of the Semantic Web and this concrete application scenario.

3.1 Performance

Since this approach has first been introduced [5], the performance of the prototype has been significantly improved. However, production of the translation phrases still takes considerable time[6], even with the very small data sets in the given scenario. Thus, this approach might be unfeasible for realtime-NLP-scenarios.

A typical answer could be the deployment of a triple store. But when following this reasoning, one could came up with the idea of using an even more optimized, application-specific storage structure and to treat LOD just as input format for a extraction, transformation and load process. Unquestionable, the LOD-vision ends with the step of transformation into a non-RDF system, when cutting off the links to the outside world - does it? Is a SPARQL-capable triple store the highest possible storage form in a properly designed LOD application?

3.2 DBpedia Endpoint

The author was thrilled to work with the real DBpedia endpoint iSPARQL[7]. Unfortunately, the designed query seems to be to expensive[8] for the endpoint policies.

While the author of course understands the technical necessity of such restrictions - is it really true that the endpoint to the world's largest pooled collection of LOD cannot be queried above a simple level complexity?

Is it best practice for LOD applications with moderate complex SPARQL queries to extract and hold LOD for its own and not working with the live DBpedia or other endpoints?

3.3 Incorporating statistics

A lot of NLP applications utilize various forms of statistics, e.g. in so called n-grams[9]. Disastrously, one basic principle of the Semantic Web makes creation of statistics hard and even counting "difficult" [1, page 252]: The Open World Assumption - the fact that "at any time [...] new information could come to light" [1, page 10]. So, it is very questionable if, for example, counting frequencies of certain triple combinations is a valid operation in the Semantic Web.

[5] http://github.com/heussd/swmt (accessed 2013-03-01).

[6] On the author's 2.4 GHz Intel Core 2 Duo machine, OWL-reasoning and quering the data takes about one second.

[7] http://dbpedia.org/isparql/ (accessed 2013-03-01).

[8] The service returns the error message: "The estimated execution time 7219 (sec) exceeds the limit of 3000 (sec)"

[9] http://en.wikipedia.org/wiki/N-gram (accessed 2013-03-02).

However, in SWMT, statistics would be required to reveal the "best" or "the most fitting" translation for a given phrase. Currently, SWMT can only find equally weighted translation alternatives - and it's pure coincidence if the phrase `Word by MS` is translated with `Word produziert von MS` or with `Word von MS`, because both forms are valid translations (see figure 2 on page 3).

3.4 Restrictions of RDF

A fundamental withdraw of the approach is the fact that it is limited to translation of triples, consisting of the three RDF elements subject, predicate and object [1, page 31]. These triples might usually resolve to small phrases of three, sometimes four words (as shown in figure 2 on page 3), but they will never constitute complete English sentences of a medium or high complexity.

The project SPARQL2NL[10], although having available a very impressive demonstration, basically seems to work similar to SWMT and thus they both underlie the same triple-restriction

However, this leads to a more general question: Does the LOD claim to carry the complete sense of a (human) language? Can any natural language sentence be converted to n triples - and can this conversion be seamlessly reversed, preserving at least the original meaning (not to speak of the exact wording)?

NLP applications usually have a language model, that includes the language's morphology, syntax and semantics. One hint could be including that knowledge of a language in the SWMT translation process. Therefore, the NLP Interchange Format (NIF)[11] looks right in doing this *the LOD way.*

[10] `http://blog.aksw.org/2013/two-aksw-papers-at-www/` (accessed 2013-02-18).

[11] `http://svn.aksw.org/papers/2012/WWW_NIF/public.pdf` (accessed 2013-03-02).

4 Conclusion

This paper describes a combinatory approach to support a Machine Translation (MT) process with world knowledge that is inherent in the Web of Linked Open Data (LOD). With a simple but expressive demo scenario, a prototype application is developed and the synergistic addition of Semantic Web technology is shown. Then, the author reflects his current stage of work and the issues he encountered in connection with application performance, DBpedia connectivity, integration of statistics and the Resource Description Framework (RDF).

Based on the feedback and the workshop results, the author continues his research in the fields of Natural Language Processing (NLP) and Semantic Web.

References

[1] Dean Allemang and James A. Hendler. *Semantic Web for the Working Ontologist - Effective Modeling in RDFS and OWL, Second Edition.* Morgan Kaufmann, 2011. ISBN 978-0-12-385965-5.

[2] Tim Berners-Lee. Linked Data. Webpage, June 2009. URL `http://www.w3.org/DesignIssues/LinkedData.html`.

[3] John G. Breslin, Alexandre Passant, and Stefan Decker. *The Social Semantic Web.* Springer, Berlin, 2009. ISBN 978-3-642-01171-9. doi: 10.1007/978-3-642-01172-6.

[4] R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. 2004.

[5] Bettina Harriehausen-Mühlbauer and Timm Heuss. Semantic Web based Machine Translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 1–9, Avignon, France, April 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W12-0101`.

[6] G. Klyne and J.J. Carroll. Resource description framework (RDF): Concepts and Abstract Syntax. Technical Report February, W3C, 2004.