Language Resources for the Semantic Web – perspectives for Machine Translation –

Cristina VERTAN
Natural Language Systems Division, University of Hamburg
Vogt-Kölln Strasse 30
22527 Hamburg, Germany
cri@nats.informatik.uni-hamburg.de

Abstract

In this paper we present a possible solution for improving the quality of on-line translation systems, using mechanisms and standards from Semantic Web. We focus on Example based machine translation and the automatization of the translation examples extraction by means of RDF-repositories.

1. Introduction

Machine Translation (MT) was nominated on the first place among the 10 emerging technologies who will change the world (Technical Review 2004). It is expected that with the increased number of official language in Europe, and the continuous growth of non-English Internet resources, machine translation systems will become an indispensable tool in everyday work. For the moment high-quality MT-systems are on one hand expensive and on the other hand domain oriented. The on-line existent tools produce poor-quality translation, and very often offer a false image of current translation engines capabilities. The main reason why on-line machine translation tools offer so poor results is that they rely either on corpus-based methods trained on a limited number of examples or they infer rules from a limited linguistic knowledge base (Gaspari 2002).

Following the statistics published in (McLaughlin and Schwall 1998) already in 1998 there were at least 25 countries with more than 500 000 Internet users, and in at least half of these countries English is neither the first nor the second spoken language. This statistic shows clearly that access to on-line information can be guaranteed only through high-quality on-line machine translation tools. However, an on-line translation system has a number of specific requirements (i.e. different from the "traditional" ones):

- It has to be fast but not always perfect. The translation of web-documents is more a kind of "translation for assimilation" in the Carbonell's classification (Carbonell 1994). However it has to go beyond the word-to-word quality offered by the actual on-line systems
- A large number of languages / pair of languages have to be covered
- The system has to be a "fully integrated black box". Most part of the users do not have the expertise to tune different parameters.

There are different approaches to automatic translation, however not all of them are suited to be used for on-line translation.

1. Rule-based MT systems are based on complex linguistic modules both in the analysis and generation phase (morphology, syntax, semantics, pragmatics). Such modules are developed for only few languages and they are not commercially –free available. The implementation of such modules requires deep linguistic knowledge in both languages (especially for the transfer rules)
2. Knowledge-based MT systems are strongly domain dependent and rely on domain-specific ontologies. Most part of the ontologies were developed previously only for commercial products, and therefore are not free available
3. Corpus-based MT systems (example – based and statistical-based) are younger on the market, and provide good translation quality, especially for assimilation purposes. They are based on large parallel aligned corpora, or on translation databases. In the first case considerable amount of text is aligned usually at the paragraph level; in the latter translation chunks are collected (usually the chunks are sentences or even smaller units.)

Most part of the currently existent on-line translation systems adopt a very simplistic rule-based approach, i.e. the translation is reduced to dictionary look-up followed by a morphological processing, and very simple syntactic transfer rules.

Within the Semantic Web activities it is assumed that a big amount of internet resources will be semantically annotated. This opens new perspectives for the corpus-based MT Systems, and makes them a serious candidate for on-line translation.

This paper is organised as follows: in section 2 we present the main principles of semantic web. In section 3 we describe a type of MT-System who can benefit from the Semantic Web activities. and show how Semantic Web technologies can be used to improve the quality of on-line Machine Translation systems. In section 4 we present directions of future work.

2. The Semantic Web

Following the definition of Tim-Berners-Lee, "The Semantic Web will bring structure to the meaningful content of the web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users"(Berners-Lee and Hendler and Lasilla 1999)

The WWW, was developed for humans; the documents on the web are machine readable but not machine understandable. The main aim of Semantic Web is to enrich documents with semantic information about the content and to develop powerful mechanisms capable of interpreting this information. These goals are achieved through implementation of models, standards as well as annotation of resources at the following layers (Berners-Lee 2003) presented in Figure 1:

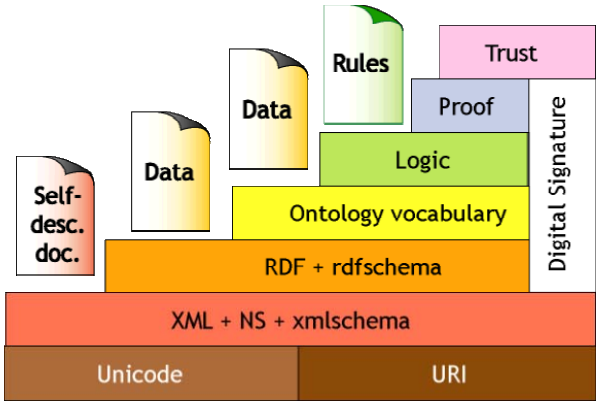

Figure 1 . Layer –cake architecture of Semantic Web (from Tim-Berners-Lee)

Unicode and URI's are the basic "bricks" in this schema, the first ensuring internationalization, the latter unique identification of any resource on the Web. XML together with its syntactic validation language XMLschema and the Name Spaces mechanism are the standard way of encoding resources. However XML tags cannot describe contents of documents. Therefore RDF (Resource Description Framework) model has to be used, and the concepts used for semantic description have to be organised in ontologies. Inference on these concepts are made at the Logic and Proof levels.

For the purposes of this article we will concentrate on the Data-levels, i.e. annotations of documents (RDF) and structure of the semantic information (Ontologies)

2.1. Document annotation with RDF

The Resource Description Framework (RDF) [is an entity relationship model used for representing information about resources in the World Wide Web. The main principle is that everything on the web can be unique identified with URI's (Uniforme Resource Identifier) and then described in terms of triples representing the resources, their properties and values. For the purposes of Semantic Web the serialization was done in XML; in this way the model benefits also from the Namespace property of XML and the RDF properties can be unique identified, independent of the users

2.2. Ontologies for Semantic Web

Ontology, a well-known Knowledge-Representation mechanism was rediscovered for the purposes of Semantic Web. The RDF properties can be organised in classes and subclasses, with attributes and values. Languages as RDFS, DAML+OIL, or recently OWL, permit complete description of complicated ontological relations between RDF properties, in an RDF/XML format. For the moment there are already hundreds of Semantic Web ontolgies for different domains, most part of them free available.

3. On-line Machine Translation and the Semantic Web

In this section we will explain first the main principles of example-based machine translation. Then we will have a closer look on how it can benefit from the Semantic Web activities.

3.1. Example-based Machine Translation (EBMT)

The basic idea in EBMT is quite simple: for the translation of a sentence previous translation examples are used. The main assumption behind this idea is that many translations are simple modifications of previous translations [CarlWay03]. In contrast with the translation memories, the selection between more possible translations is completely automatic.

A typical EBMT System is based on the following components (Trujillo 1999)

1. A database of aligned sentences in the source and target languages. The contents of the database, as well as its dimension are essential for the quality of the selection. The examples have to be domain-relevant, long enough to capture specific particularities of a construction and short enough to be retrieved in common texts
2. A matching algorithm that identifies the examples that most closely resemble all or part of the input sentence
3. A combination algorithm which rebuilds the input sentence, through a combination of retrieved fragments
4. A transfer and composition algorithm that extracts corresponding target fragments and combines them into a sentence in the target language.

It turned out that information about the syntactic structure of the fragments in both languages as well as pattern transfer rules, can improve significantly the performance of the example-based MT system.[Carlway03]. Therefore it is quite usual that the example database contains, together with parallel aligned strings, also syntactic structures and their correspondences.

3.2. Language Resources for Semantic Web and their role in Machine Translation

Between the main activities in the Semantic Web at the moment we encounter:

- the description and annotation of a large number of web resources following the RDF model
- the creation of repositories of RDF properties, organised in ontologies.

Every resource (document piece of document or even sentence) is described via a triple (Subject, Predicate, Object). All three elements of the triple refer to the logical structure of the resource and not the syntactic one. It is expected that in the near future, a big part of the documents in Internet will be annotated following the RDF model.

Machine Translation, and in particular Example-based Machine Translation can make use of these additional annotations for three purposes:

1. For the achievement of parallel aligned corpora. Small languages still suffer from lack of linguistic resources, and especially multilingual resources. On-line documents are main source for machine-readable corpora, however, with few exceptions (explicitly translations of the same Web page) it is difficult to determine automatically which part of a document is a translation of another document. RDF annotations can be used for such purposes
2. For Example based rough translation: As mentioned in section 1 on-line translation is made for assimilation purposes, therefore, meaning preservation is much more important as an exact translation. RDF model aims to enrich documents with information about their content. This can help in the process of "example based rough translation". Until now, the trials in this field were done only on the basis of retrieval and translation of content-words [ShimhataSumitaMatsumoto03].
3. For disambiguation: the current example based translation systems make use only of syntactic annotation. These can be insufficient in disambiguation cases like the following:

Let us assume that we have in the database of translation examples:

Große Besonderheiten ↔ important peculiarities

Große Städte ↔ big cities

The translation choice for große Schlößer as important castles or big castles is context depending. For the moment the disambiguation is done only statistical. Semantic annotation of the examples , as well as the input text would increase the translation accuracy. This makes sense especially for translation of on-line resources which are supposed to be correspondingly annotated

Although the advantages of Semantic Web annotations (in particular RDF-model) are transparent from the points mentioned above, the main question which arises is

Who will decide which semantic information has to be included, at what level (sentence

/paragraph/document), and in which language?

Following information is needed for increasing the translation quality :

- translation equivalents of words /expressions
- transfer rules for syntactic structures
- semantic classes for the candidate solutions.

The main problem to be solved is the consistency between different RDF annotations corresponding to different users. Let us assume that in the German text the annotation for Große Städte is .

<rdf.description rdf. about:"http…..>
<user1: Messung > Große </user1: Messung >

and in the English one
<rdf.description rdf. about:"http…..>
<user2: size >big</user2: size >

A relationship between "size" and "Messung" has to be established showing that they refer to the same concept. This has to be done via mapping on an ontology. The main challenge in the design of ontologies with multilingual instances is that, very often words in one language overlap concepts in the ontology, and there is no one-to-one mapping to the meaning in the other language

The architecture in figure 2 proposes a framework for extracting translation correspondences, taking into account their RDF annotations. We propose the organisation of the RDF annotation scheme in two parts: syntactic annotation and semantic annotation. The concepts to be instantiated for this annotations will be organised in two correspondent ontologies.

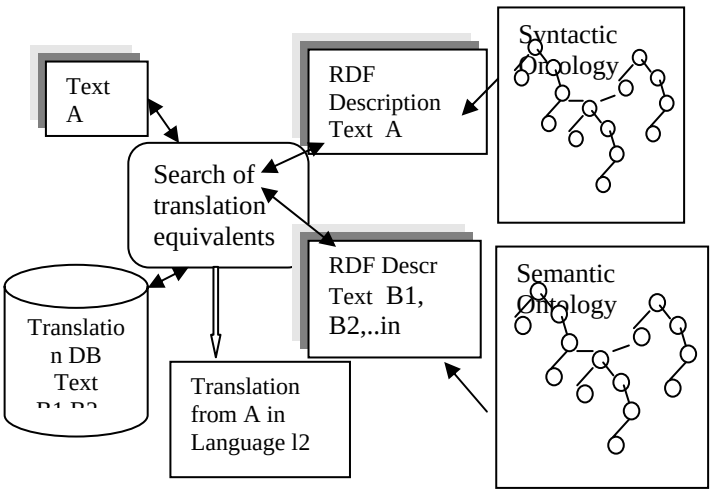

Figure 2: Extraction of Translation Equivalents from RDF annotated texts.

Assuming that input is a text A in language L1, a search process will identify fragments from A in the translation database and obtain one or more translations, namely Texts B1, B2,…Bn. During the next step the RDF descriptions of the input text and the translation candidates are compared by mapping the RDF annotations on the syntactic and semantic ontology, and the most similar one is chosen as output.

At the University of Hamburg we are currently implementing this schema within a Demo-System for German and English texts, in tourist domain. Approximately 30 documents in both languages are currently annotated with linguistic properties in RDF format, mapped on a syntactic respectively semantic ontology.

4. Conclusions and Further Work

In this article we presented the main principles of semantic Web as well as its possible contributions to the improvement of on-line translation systems. A solution for automatic extraction of translation examples from RDF-annotated texts is also presented. However the architecture supposes the existence of the repositories for syntactic and semantic annotations as well as the both ontologies. In order to ensure the viability of the principle for on-line translation systems, such repositories have to be created for different languages, texts have to be annotated and the ontologies have to cover a broad spectrum of linguistic phenomena.

After the complete implementation of the demo system we intend to perform an evaluation of the translation quality, and to analyse also the accuracy of the extraction mechanism,

References

T. Berners-Lee 2003, Foreword to "Spinning the Semantic Web-Bringing the World wide Web t Its Full Potential", in D. Fensel, D., J. Hendler,, H.Lieberman, and W. Wahlster, (eds.), MIT Press, 2003

T. Berners-Lee, and J. Hendler, and O. Lasilla, 1999, "The Semantic Web", Scientific American, 1999

J. Carbonell 1994, Slides of a tutorial on MT Saarbrücken 1994. unpublished

A-Way, and M. Carl 2003, "Introduction to Example-based machine Translation", Kluwer Academic Press, 2003

F. Gaspari 2002 "Using free on-line services in MT teaching," in Proceedings of the 6th EAMT Workshop on Teaching Machine Translation, November 14-15, 2002, Manchester, pp.145-153

S. McLaughli, and U. Schwall 1998, “Machine Translation and the Information Soup:”, in Third Conference of the Association for Machine Translation in the Americas, Proceedings of the AMTA'98, LNAI 1529, D. Farwell et. Al. (Eds.) Langhorne, PA, USA, October 1998, pp. 384-397

M. Shimohata, and E Sumita, and Y. Matsumoto 2003, “Retrieving Meaning-equivalent Sentences for Example-based Rough Translation”, HLT-NAACL Workshop: Building and using Parallel Texts. Data Driven Machine Translation and Beyond, Edmonton, May-June 2003, pp. 50-56

Technology Review 2003, „10 emerging Technologies who will change the World”, retrieved at http://www.technologyreview.com/

A Trujillo 1999, Translation Engines: Techniques for Machine Translation, Springer Verlag, 1999