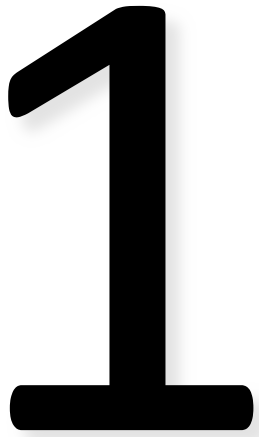
Semantic Web based Machine Translation

Bettina Harriehausen-Mühlbauer
bettina.harriehausen@h-da.de
University of Applied Sciences
Darmstadt, Germany

Timm Heuss
Timm.Heuss@web.de
University of Applied Sciences
Darmstadt, Germany

EACL Workshop ESIRMT-HyTra 2012

Agenda



1. Overview & Introduction
2. Motivation
3. Approach & Prototype
4. A critical view
5. Outlook

Overview

- We present an experimental combination of traditional **Natural Language Processing (NLP)** technology with the **W3C Semantic Web** building stack

- We provide an sample scenario in which Semantic Web technology extends the expert knowledge required for a **Machine Translation (MT)** task

Introduction

The World Wide Web

- The Web was invented by Tim Berners-Lee in 1990 at the CERN laboratories.
- Fundamental properties:
  - Decentralization
  - Interoperability
  - Openness and ease of use

Source: http://www.britannica.com/EBchecked/topic/62493/Sir-Tim-Berners-Lee

Introduction

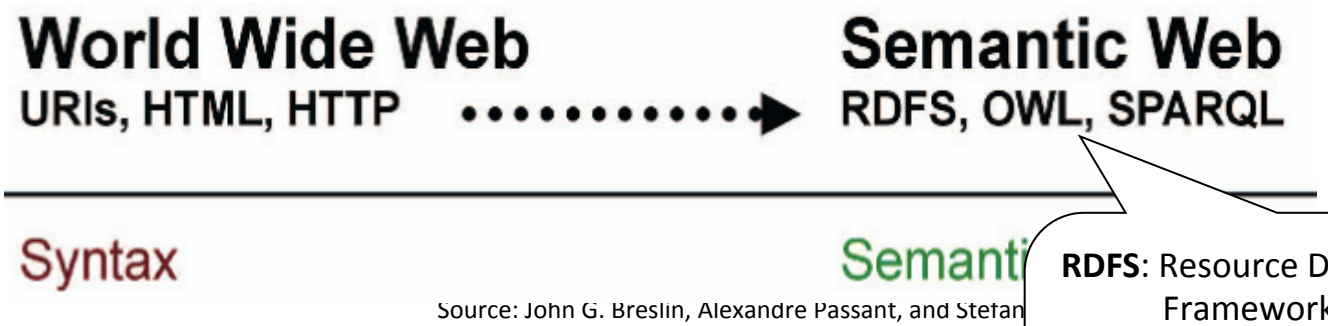
Semantic Web

Source: John G. Breslin, Alexandre Passant, and Stefan Decker: The Social Semantic Web.

- Introduced in the new millenium
- Standardized technology stack to make information available to machines

Introduction

Semantic Web



- Introduced in the new millenium
- Standardized technology stack to make information available to machines

Introduction

Paradigms in the Semantic Web

- The „AAA slogan“:

  Anyone can say Anything about Any topic

  (Allemang & Hendler)

- Open World Assumption:
  - there is always more knowledge
  - new knowledge can always be added later.

Introduction

Semantic Web basics

- RDF expresses meaning by encoding it in sets of triples / resources:

```
:subject :predicate :object .
```

(N3 notation)

- Built-in capacities for
  - a human-readable version of a resource's name (**`rdfs:label`**)
  - Query language SPARQL
  - Drawing conclusions from a knowledge base

Agenda

2

1. Overview & Introduction
2. Motivation
3. Approach & Prototype
4. A critical view
5. Outlook

Motivation

Ambiguity is a common challenge to NLP

- The problem in many areas of NLP is the **ambiguity** of natural language on various levels, from word level to sentence level.
- In many cases, strings can only be disambiguated on the basis of **world or expert knowledge**

- Traditional solution in translation systems: **dedicated expert dictionaries**

Motivation

Expert dictionaries - A typical solution

NLP /
Expert dictionaries

- Expensive
- Hard to build
- Domain specific

Agenda

3

1. Overview & Introduction
2. Motivation
3. Approach & Prototype
4. A critical view
5. Outlook

Approach

Expert dictionaries vs. RDF

NLP /
Expert dictionaries

- Expensive
- Hard to build
- Domain specific

Semantic Web /
RDF

- Distributed, in the spirit of the Web
- Universal format
- Suiteable for a lot of application areas

Approach

Basic idea

- We believe that modern Information Technology is aligned and committed to information and its markup
- We use the contained knowledge in the Web for a disambiguation process - without additional MT rules or statistics being applied
- We benefit from central building paradigms of the Semantic Web

Approach

Expert dictionaries vs. RDF

NLP /
Expert dictionaries

- Expensive
- Hard to build
- Domain specific

Semantic Web /
RDF

- Distributed, in the spirit of the Web
- Universal format
- Suiteable for a lot of application areas

Approach

Expert dictionaries vs. RDF

NLP / Expert dictionaries

Use the Semantic Web as expert dictionary!

Semantic Web / RDF

- Distributed, in the spirit of the Web
- Universal format
- Suiteable for a lot of application areas

Prototype

Sample scenario

- We have chosen an **ambigious machine translation** from English to German to test our approach

- Sentence (English):
  Pages by Apple is better than Word by MS.

- Manual translation (German):
  Pages von Apple ist besser als Word von MS.

- Involved expert knowledge
  - A product named Pages is produced by a vendor named Apple
  - A product name Word is produced by a vendor named MS

Prototype

Semantic Web knowledge base

- Objects in the knowledge base

```
:microsoft a :vendor, :trigger .
:microsoft rdfs:label "MS" .
:microsoft :produces :word , :excel, :windows.

:apple a :vendor, :trigger .
:apple rdfs:label "Apple" .
:apple :produces :pages , :iphone .
```

- Relationships between Objects:

```
:producedby rdfs:label "by"@en,"von"@de.
:produces owl:inverseOf :producedby .
```

Prototype

Sample scenario

- We have chosen an ambigious machine translation from English to German to test our approach

- Sentence:
  Pages by Apple is better than Word by MS.

- Manual translation:
  Pages von Apple ist besser als

- Involved expert knowledge
  - **A product named Pages is produced by vendor named Apple**
  - **A product name Word is produced by a vendor named MS**

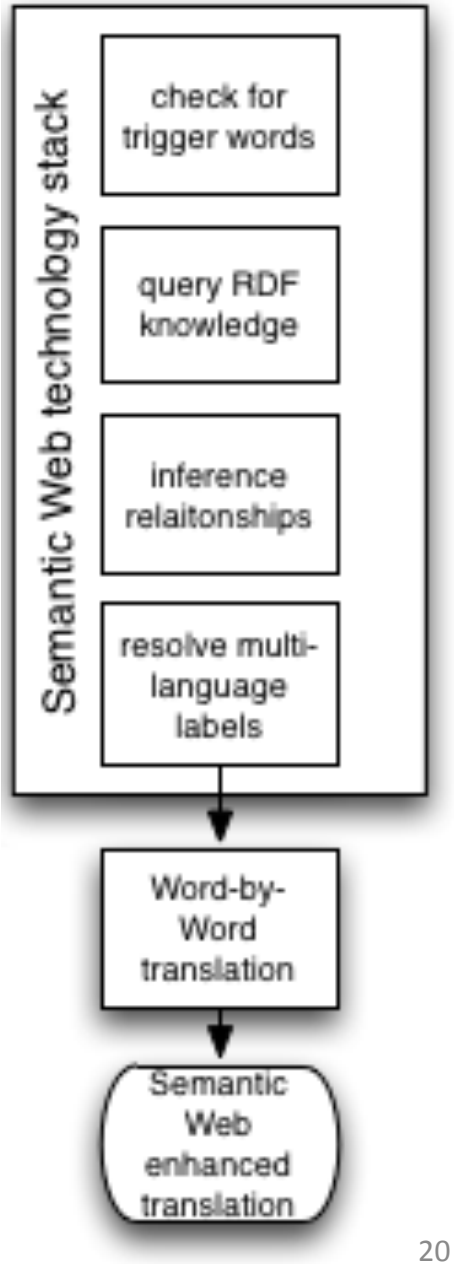
This knowledge is not explicitly defined!

Prototype

Stepwise refinement

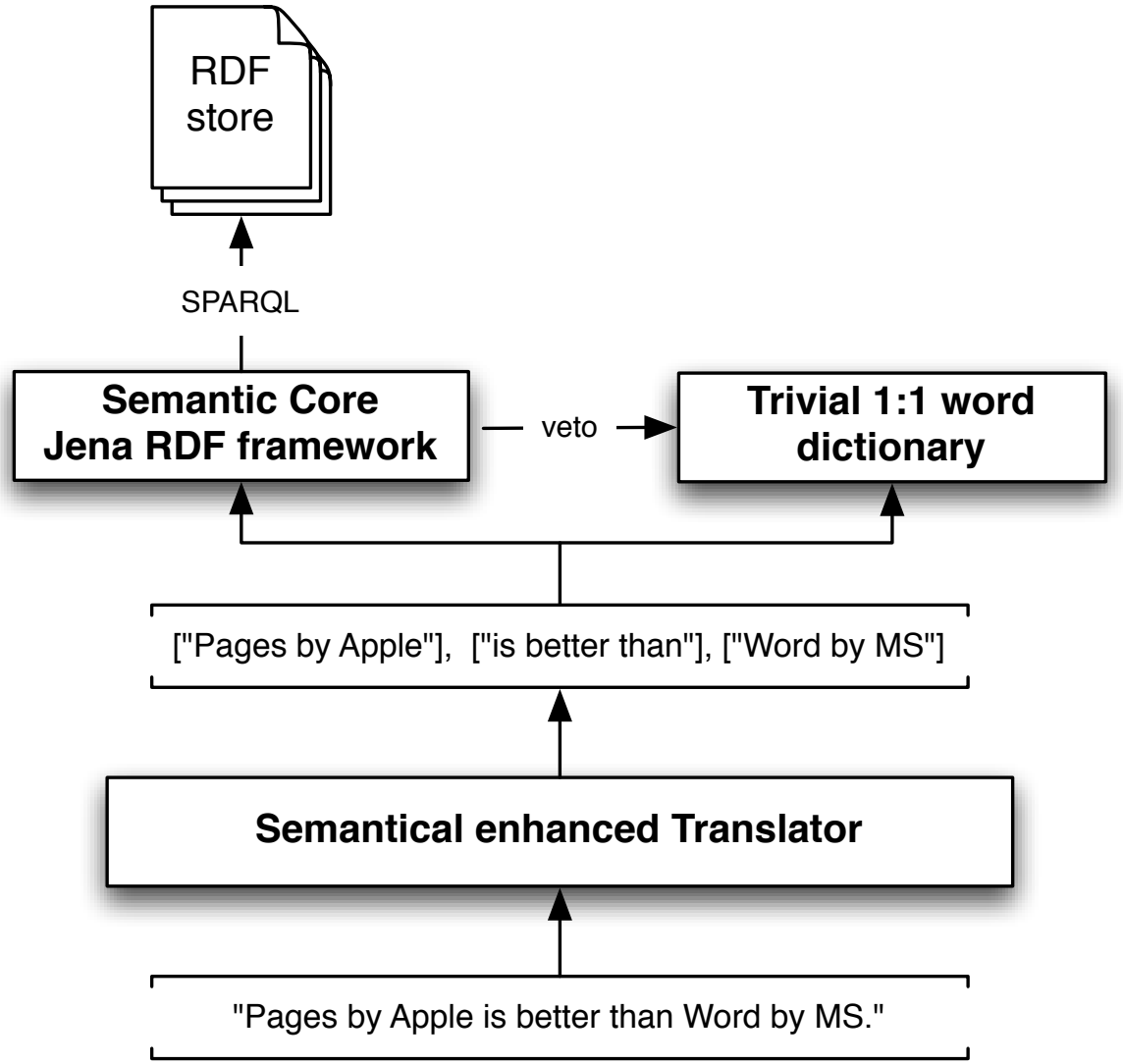
- Steps of a SWMT process
  1. Check for trigger words
  2. Query RDF knowledge
  3. Inference relationships
  4. Resolve multilanguage labels

  5. Word-by-word translation



Prototype

Component overview



Agenda

1. Overview & Introduction
2. Motivation
3. Approach & Prototype
4. A critical view
5. Outlook

A critical view

- We stand at the very beginning of our research and have encountered corresponding issues

A critical view

Performance issues

- Implementation of the program logic turned out to be quite complicated - especially the query mechanism
- A lot of SPARQL queries are fired on execution
- Reasoning takes additional time

➢ Overall performance is currently bad!

➢ Depending on the application: It needs to be discussed if this is a real disadvantage

A critical view

Usage of RDF

- The current implementation works internally with RDF triples.
- This only allows translation of simple phrases; complex sentences do currently not work with our prototype

➢ More sophisticated ways to store natural language in RDF need to be evaluated

A critical view

No golden standard

- Because of the Open World Assumption, there is no golden truth in the Semantic Web-world (because there could always lack something important)
- Getting the „best“ translation for a certain sentence just by relying on the knowledge base is not possible
- Probably, our approach does not hold for providing complete translation solutions - but maybe at least for giving qualified suggestions.

➢ A combination of established NLP techniques and our approach might be a gainful next step.

Agenda

5

1. Overview & Introduction
2. Motivation
3. Approach & Prototype
4. A critical view
5. Outlook

Outlook

- At this point in our research, we have not yet combined existing MT technology, especially Statistical MT, with SWMT.
- Existing MT technologies and SWMT are certainly not mutually exclusive and we suspect that a combination of MT approaches will lead to yet even better results, especially in cases where the translation quality is based on world or expert knowledge.

Thanks for your attention!

What do you think?

Contact me:

Timm.Heuss@web.de

http://heussd.github.com/swmt/