Evaluation of Semantic Service Discovery

A Survey and Directions for Future Research

Ulrich Küster¹, Holger Lausen², Birgitta König-Ries¹

(¹University Jena, Germany, ²DERI Innsbruck, Austria)
Agenda

- Introduction / Motivation
- IR versus SWS Evaluation – General thoughts
- State of the Art
  - S3 Matchmaker Contest
  - OWL-S Test Collection 2
  - Semantic Web Service Challenge
  - DIANE Evaluation
- Directions for Future Work
Bringing the web to its full potential...

Dynamic

Web Services
UDDI, WSDL, SOAP

Semantic Web Services

Static

WWW
URI, HTML, HTTP

Semantic Web
RDF, RDF(S), OWL

Syntactic

Semantic

(gratefully borrowed from WSMO SWS Tutorial at ICWE 2005)
Motivation of Work

- SWS research attracts much effort and money:
  - sixth EU framework: > 20 projects, > 70.000.000 €
- Many approaches, many results, many papers
- Little effort / results regarding practical evaluation
- Poor quality of evaluation

"There are many claims for such technologies in academic workshops and conferences. However, there is no scientific method of comparing the actual functionalities claimed."

W3C SWS Testbed Incubator Group Charter
"... there has been no concerted effort by groups to work with the same data, use the same evaluation techniques, and generally compare results across systems. ... Evaluation using the small collections currently available ... certainly does not demonstrate any proven abilities of these systems to operate in real-world ... environments. This is a major barrier to the transfer of these laboratory systems into the commercial world."

Donna Harman 1992
Agenda

- Introduction / Motivation
- **IR versus SWS Evaluation – General thoughts**
- State of the Art
  - S3 Matchmaker Contest
  - OWL-S Test Collection 2
  - Semantic Web Service Challenge
  - DIANE Evaluation
- Directions for Future Work
The general problem: IR versus SWS Evaluation

- **Real World**
  - Information Need
  - Abstraction
  - Model
  - Retrieval
  - Result

- **Information Supply**
Approach 1: "TREC"-like

- Evaluation based on test collection:
  - set of SWS descriptions
  - set of example requests
  - set of relevance judgments
  - process automated
  - evaluation in terms of recall, precision, ...

- Problems:
  - requires high quality test collection
  - does not evaluate description formalism
  - R/P based on semantic descriptions?
Approach 2: Scenario-based

- Evaluation based on scenarios / use cases
  - services/requests not given in particular formalism
  - modeling part of the evaluation
  - human test subjects involved
  - different evaluation measures

- Problems
  - evaluation influenced by expertise of test subjects
  - evaluation cannot be automated
  - high effort, doesn't scale well
  - issue of decoupling (offers and requests should not be modeled by same person)
AGENDA

- Introduction / Motivation
- IR versus SWS Evaluation – General thoughts
- **State of the Art**
  - S3 Matchmaker Contest
  - OWL-S Test Collection 2
  - Semantic Web Service Challenge
  - DIANE Evaluation
- Directions for Future Work
State of the Art 1: S3 Matchmaker Contest

- TREC-like
- First edition at ISWC 2007 in Busan, Korea
- Based on OWL-S Test Collection 2

- Currently applicable only to OWL-S matchmakers
- Does not evaluate OWL-S or OWL-S based modeling
- False positives/negatives to be attributed to matchmaker or descriptions?
  e.g. Klusch & Fries 2007: "Hybrid OWL-S Service Retrieval with OWLS-MX: Benefits and Pitfalls"
- Flaws of OWLS-TC2
State of the Art 2: OWLS-TC2

- OWL-S Test Collection 2 (http://projects.semwebcentral.org/projects/owls-tc)
- Currently biggest (and best?) public SWS collection
- Many problems
  - many unrealistic services (decreases real-world relevance)
  - semantically poorly described services
    - only IO, flat concepts only, no attributes, no relations
    - e.g.: input "Car", output "Prize", "Auto" (subclass of "Car")
  - offers reverse engineered from requests (e.g. 11 services that offer package of car + bycicle)
  - collection developed for particular, hybrid matchmaker

- Context
  - developed by single group! (Klusch et al.)
  - community contribution encouraged and required: http://www-ags.dfki.uni-sb.de/swstc-wiki
State of the Art 3: SWS Challenge

- http://sws-challenge.org
- W3C Incubator Activity: http://www.w3.org/2005/Incubator/swsc/
- Based on 3 scenarios (mediation and discovery)
- Design principles:
  - solution independence (don't reverse engineer evaluation)
  - language neutral (best formalism core research question)
  - no participation without invocation (backed by implementation)
- Evaluation of functional coverage + solution flexibility
- Peer code review of solutions + changes in scenarios
- Problems:
  - so far no objective measurement of flexibility / effort
  - large effort due to implementation overhead
  - small set of services
  - disregards issue of decoupling
State of the Art 4: DIANE Evaluation

- Project on automation of service usage
- Evaluation of expressiveness of formalism, decoupling, degree of automation and efficiency of matchmaking
  http://hnsp.inf-bb.uni-jena.de/DIANE/benchmark
- Expressiveness
  - Created set of "real-world" services by asking volunteers (three scenarios, 200 service requests)
  - evaluation by ability to encode requests
- Decoupling
  - same services independently encoded as offers and requests
  - measurement of recall and precision of matchmaking
- Degree of automation + efficiency
  - proof of concept implementation

- Remarks
  - surprisingly difficult to come up with services
  - results vary greatly among scenarios
  - expressiveness rating not well funded
  - benchmark so far never applied to another approach
Results

- A) Surprisingly little effort devoted to evaluation
  - in most projects no evaluation to speak (or not published)
  - BUT recently increase in attention
- B) Lack of theory in SWS evaluation
  - no solid discussion of scope, dimensions, measures, ...
  - no meta-evaluation
- C) Existing approaches only starting points
  - even though promising, all with flaws
  - complementary but isolated
Agenda

- Introduction / Motivation
- IR versus SWS Evaluation – General thoughts
- State of the Art
  - S3 Matchmaker Contest
  - OWL-S Test Collection 2
  - Semantic Web Service Challenge
  - DIANE Evaluation
- Directions for Future Work
Directions for Future Work 1

- Towards a standard SWS evaluation methodology
  - what to evaluate? ➔ criteria (e.g. expressivity)
  - how to evaluate? ➔ measures (e.g. functional coverage of use cases)
  - how to measure? ➔ measuring instruments (e.g. collection of use cases)
  - how to implement? ➔ methodologies (e.g. experimental setup)

- how to achieve validity, reliability, efficiency?
- requirements towards evaluation, criteria for meta-evaluation
"It's not that no metric is available. It's that the suggested metrics all have obvious deficiencies, none are widely used, and that there is relatively little discussion about how to improve them."

Directions for Future Work 2

- Towards a common SWS test bed
  - work towards unifying approaches
  - develop more effective means of sharing experiences, results and test data

- Ongoing work: Common SWS portal
  - current collections scattered, often private
  - usually collection of flat files
  - poorly documented
  - poorly structured
  - no support for search, rating, easy editing, ...
Suggested SWS Collection structure

WordNet

I know what you meant...

Service described in natural language

IO declared in natural language

Categorization

WSDL descriptions

Semantic descriptions (OWL-S, WSML, SAWSDL, …)

Tagging

Thanks You!

Questions?

Ulrich Küster (Ulrich.Kuester@uni-jena.de)
http://hnsp.inf-bb.uni-jena.de/ukuester