Ontologies for Crosslingual Applications

MT Summit X Workshop:
Semantic Web Technologies for Machine Translation

12 September 2005

Semantic Web Technologies, Crosslingual Applications, & Hybrid Approaches to MT

MT Summit X Workshop:
Semantic Web Technologies for Machine Translation

12 September 2005

Why do I think it‘s time again for knowledge-based MT?

MT Summit X Workshop:
Semantic Web Technologies for Machine Translation

12 September 2005

OUTLINE

- ☆ Semantic Web and Multilingualism
- ☆ Crosslingual Technologies
- ☆ Ontologies for Crosslingual Technologies
- ☆ Obstacles for the Semantic Web
- ☆ Two Perspectives
- ☆ MT: Major Obstacles
- ☆ MT: Exciting New Developments
- ☆ MT: Hybrid Machine Translation Approaches
- ☆ Open Source Systems for Hybrid MT

- ☆ Semantic Web as a solution to the technological challenges of multilingualism
- ☆ Semantic Web as a challenge for the technological solutions to multilingualism

- ☆ Faulty claim: The multilingual setup will become less important because the core of human knowledge will be stored in a language-independent way

- ☆ Still needed dictionaries (and terminologies) that link words in human languages to the “language independent” representation of concepts

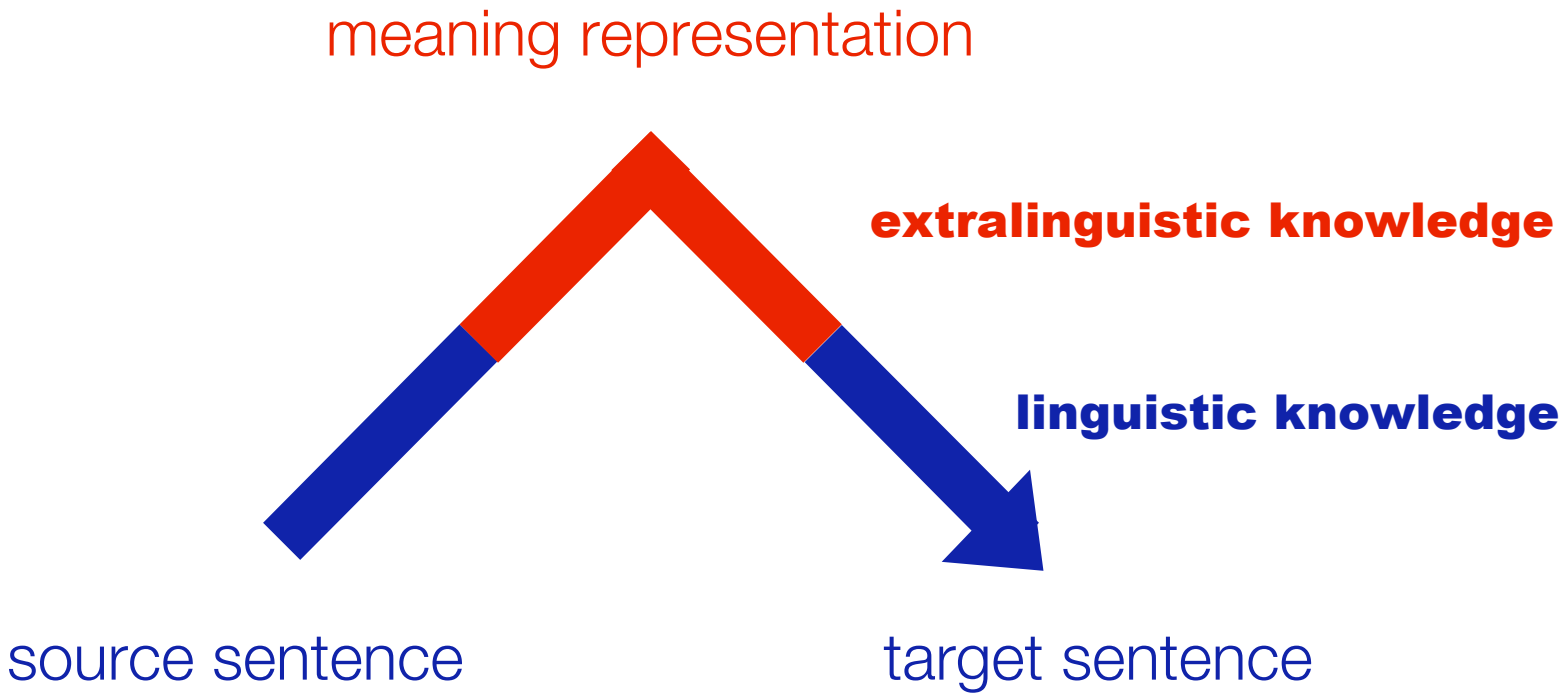
- ☆ Crosslingual Information Retrieval (CLIR)
- ☆ Crosslingual Summarization
- ☆ Crosslingual Information Extraction (actually IE plus Multilingual Generation)
- ☆ Crosslingual Question Answering
- ☆ Machine Translation

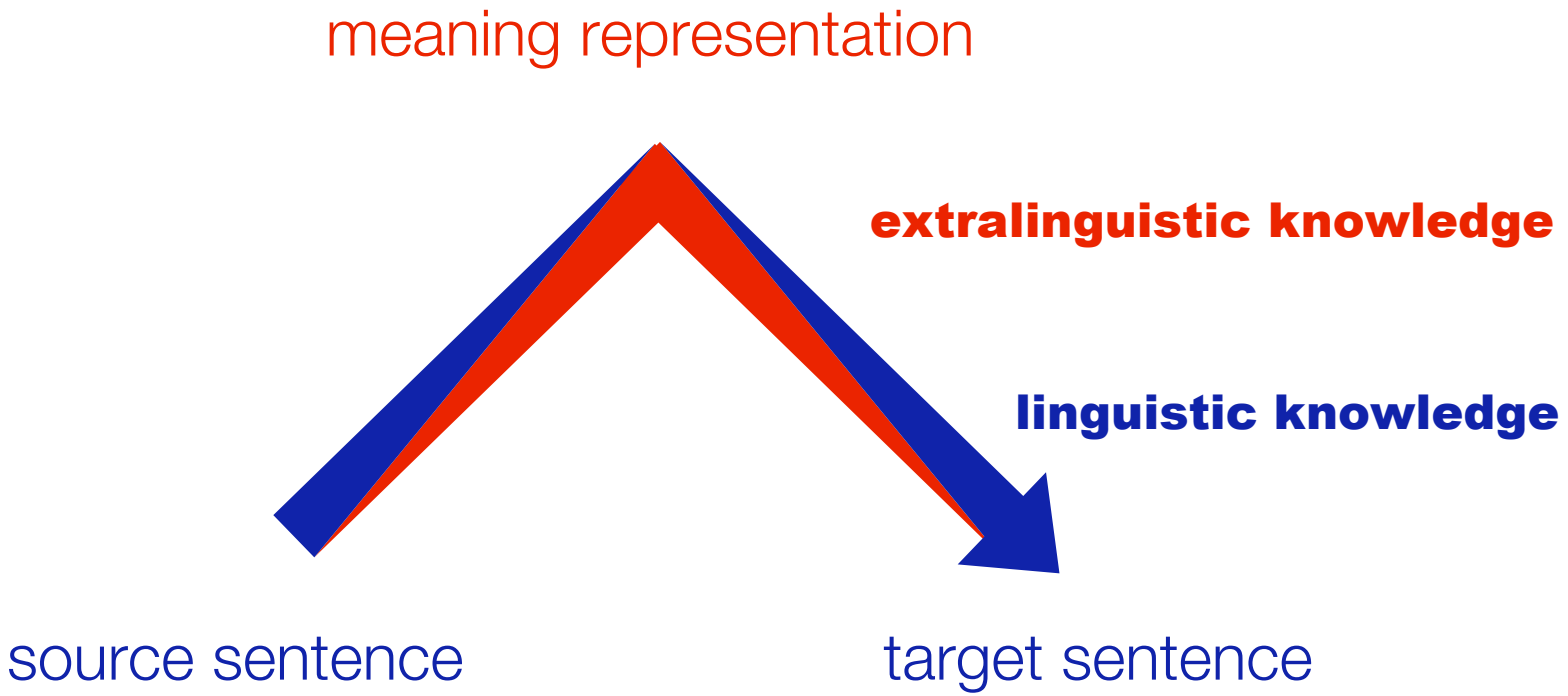
- ☆ document translation for indexing
- ☆ query translation
- ☆ document translation for returned documents
- ☆ no MT but crosslingual indexing

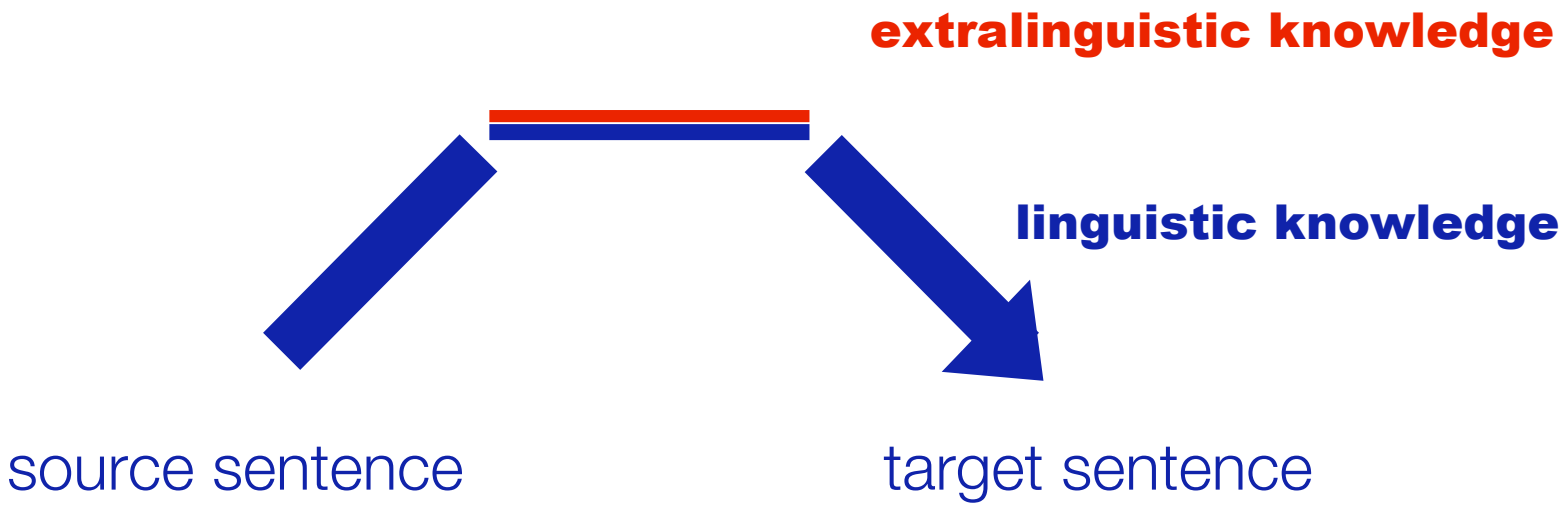
- ☆ use of taxonomies and ontologies for
- ☆ example: project MUCHMORE
  - medical ontologies UMLS, SNOWMED, MeSH
  - combination with statistical methods

Uses of Knowledge in MT

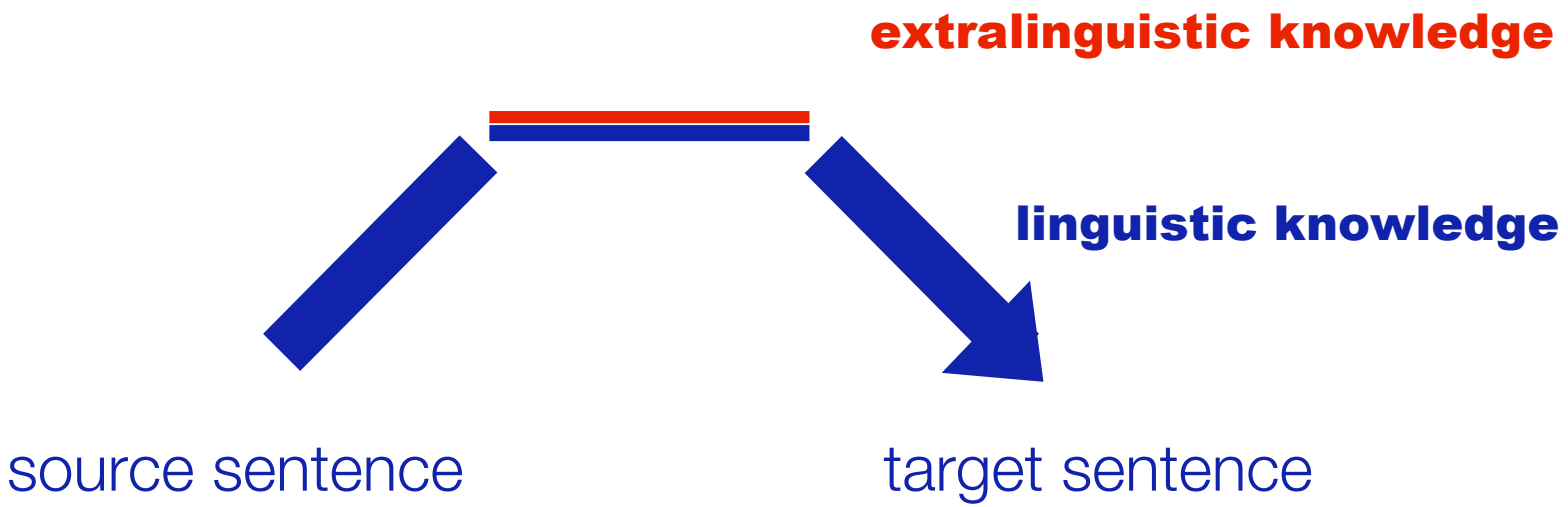
- ☆ MT based on linguistic knowledge: linguistic or rule-based MT
- ☆ MT based on explicit representation of non-linguistic knowledge: **KBMT**
- ☆ **MT based on implicit representation of linguistic and nonlinguistic knowledge: non-statistical EBMT and statistical MT**







- ☆ outdated solutions to linguistic representation & processing
- ☆ outdated solutions to representing extralinguistic knowledge



implicit
linguistic and
extralinguistic
knowledge

source sentence

target sentence

☆ Past decades: No progress for more than 25 years

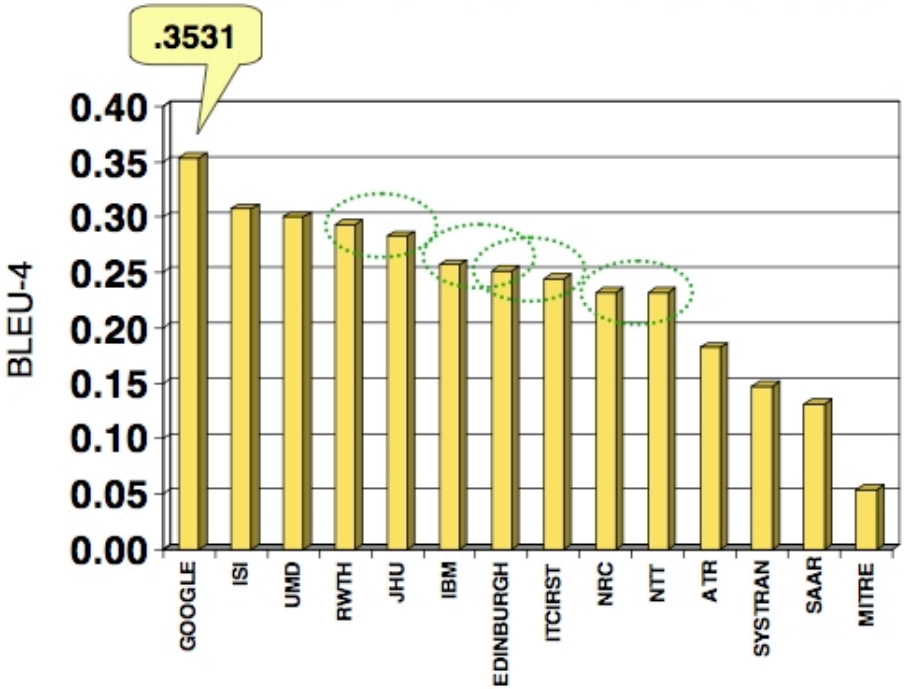
☆ Today: Strong reasons for cautious optimism

Time for Progress

- ☆ Use of vast volumes of parallel data on the web (for a few language pairs) has improved statistical translation at least for those languages.
- ☆ New learning methods such as co-training and active learning have emerged that will greatly reduce the volume of needed training data.
- ☆ Due to progress in algorithms and hardware, deep linguistic processing has become efficient. Grammar engineering has become affordable.
- ☆ The most promising development is the combination of the improved statistical methods with the improved knowledge-driven methods in a variety of clever ways.

Results – Chinese Large Data Track

Primary systems ordered by overall test set score



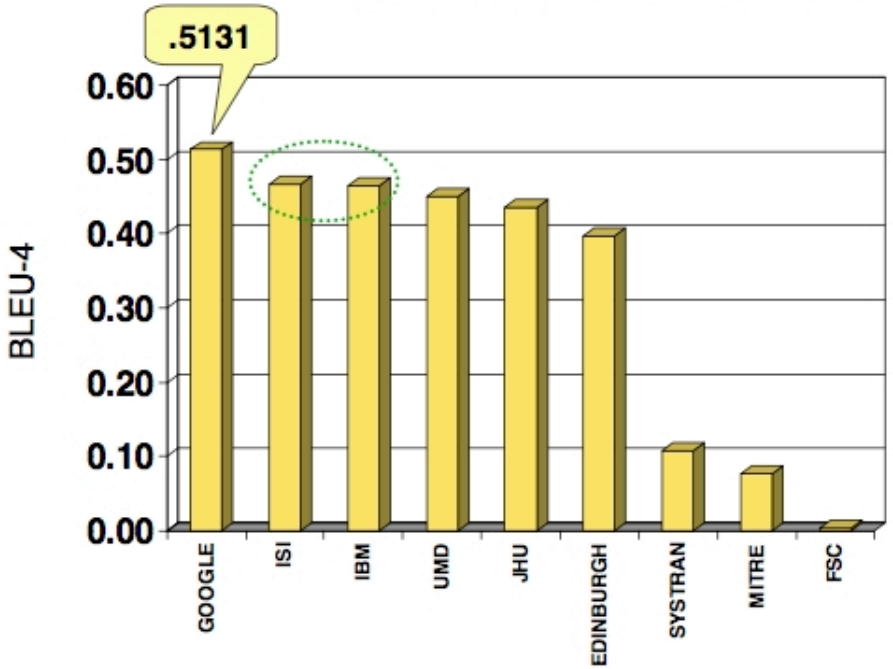
- Oval indicates not enough evidence to reject the null hypothesis that site1 and site2 are **equally likely** to receive a higher BLEU score on any given document

NIST MT-05 Evaluation Workshop

NIST National Institute of Standards and Technology

Results – Arabic Large Data Track

Primary systems ordered by overall test set score



- Oval indicates not enough evidence to reject the null hypothesis that site1 and site2 are **equally likely** to receive a higher BLEU score on any given document

NIST MT-05 Evaluation Workshop

NIST National Institute of Standards and Technology

Examples:

- MT for indexing and indicative translation
- treebanks and statistical MT
- information extraction and MT
- statistical methods for the choice among alternatives in linguistic systems

- ☆ a statistical system learns from data
- ☆ during translation, it notices gaps in "acquired knowledge"
- ☆ system then demands new data
- ☆ needed data can be greatly reduced
- ☆ chances for language pairs without large parallel corpora
- ☆ over-learning effects can be avoided

- ☆ data request is now formulated as set of sentences in source language
- ☆ idea # 1: let the machine try to translate the sentence by using related words extracted from WordNet, pose alternatives to the human translator
- ☆ idea # 2: use learning and active learning for a combination of linguistic and corpus-based methods

☆ Progress especially in two communities:

LFG (PARGRAM)

HPSG (DELPH-IN)

DELPH-IN

- ☆ Cambridge University (UK), Computer Laboratory
- ☆ DFKI Saarbrücken GmbH (Germany), Language Technology Lab (co-founder)
- ☆ NTT Communication Science Laboratory (Japan), MT Research Group
- ☆ Norwegian University of Science and Technology (Norway), Lingvistisk Institutt
- ☆ Saarland University (Germany), Department for Computational Linguistics
- ☆ Stanford University (US), LinGO Laboratory at CSLI (co-founder)
- ☆ Tokyo University (Japan), Tsujii Laboratory
- ☆ University of Oslo (Norway), MT Research Group
- ☆ University of Sussex (UK), School of Cognitive and Computing Sciences
- ☆ University of Washington (US), Computational Linguistics Laboratory
- ☆ *University of Lisbon*
- ☆ *University Pompeu Fabra Barcelona*
- ☆ *NCSR Demokritos*
- ☆ *Kyung Hee University Seoul*
- ☆ *CNRS LORIA*
- ☆ *University Linköping*

Components of DELPH-IN

- ☆ Joint Computational Formalism (set by ERG Grammar and LKB)
- ☆ Grammar Development Tools (LKB)
- ☆ An Interlingual Core Grammar (The Matrix)
- ☆ Implemented Grammars (ERG, Japan., French, German, Greek,...)
- ☆ HPSG Treebanks (Redwoods, Eiche, Hinoki)
- ☆ Parsers (PET, LILFES, ...)
- ☆ Generator (in the LKB)
- ☆ Engineering Platform (tsdb)
- ☆ Platform for Hybrid Processing (HoG)
- ☆ Comparative Evaluations
- ☆ Exchange, Cooperation and Mutual Assistance
- ☆ Joint Promotion and Project Acquisition

- ☆ Heart of Gold (HoG) Platform for hybrid processing
- ☆ (Robust) Minimal Recursion Semantics (R) MRS Common semantic formalism
- ☆ Active research on combination with ontologie exploitation
  - Cambridge
  - Edinburgh
  - Saarbruecken

- ☆ Really-existing Web: Gigantic Web
- ☆ Powerful Vision of a Future Web: Semantic Web

- ☆ Gigantic Web
  - catalogued: 15 billion pages
  - estimated tital size: 550 billion pages

- ☆ Semantic Web
  - Several hundred ontologies (number keeps growing)
  - most of them small
  - some are connected with large numbers of instances
    such as SWETO with approx. 1 Mio instances and 1.5 Mio relations

- ☆ No connection between Gigantic Web and Semantic Web
- ☆ One promising connection:
  - learning of ontologies with instances from texts
- ☆ Another promising connection:
  - gradual enrichment of web contents with multilayer standoff markup
- ☆ These two developments can go together

- **Buitelaar and Sintek 2004: OntoLT Version 1.0: Middleware for Ontology Extraction from Text**

  Protégé PlugIn with connection to the IE component SHUG (Declerck)

☆ Buitelaar, Eigner, Declerck (2004) **OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection**

☆ **OntoSelect: Towards the Integration of an Ontology Library, Ontology Selection and Knowledge Markup**
Proceedings of the Workshop on Knowledge Markup and Semantic Annotation (Semannot2004) at the International Semantic Web Conference, Hiroshima, Japan November 2004

- ☆ Specialized MT for selected purposes:
  - weather reports
  - avalanche warnings
  - Caterpillar manuals
  - EU documentation
- ☆ Sometimes such specialized systems need to be combined
- ☆ Example from our own wor: COMPASS 2008

- ☆ Multilingual, Mobile, Multimodal Information Services for the 2008 Beijing Olympic Games

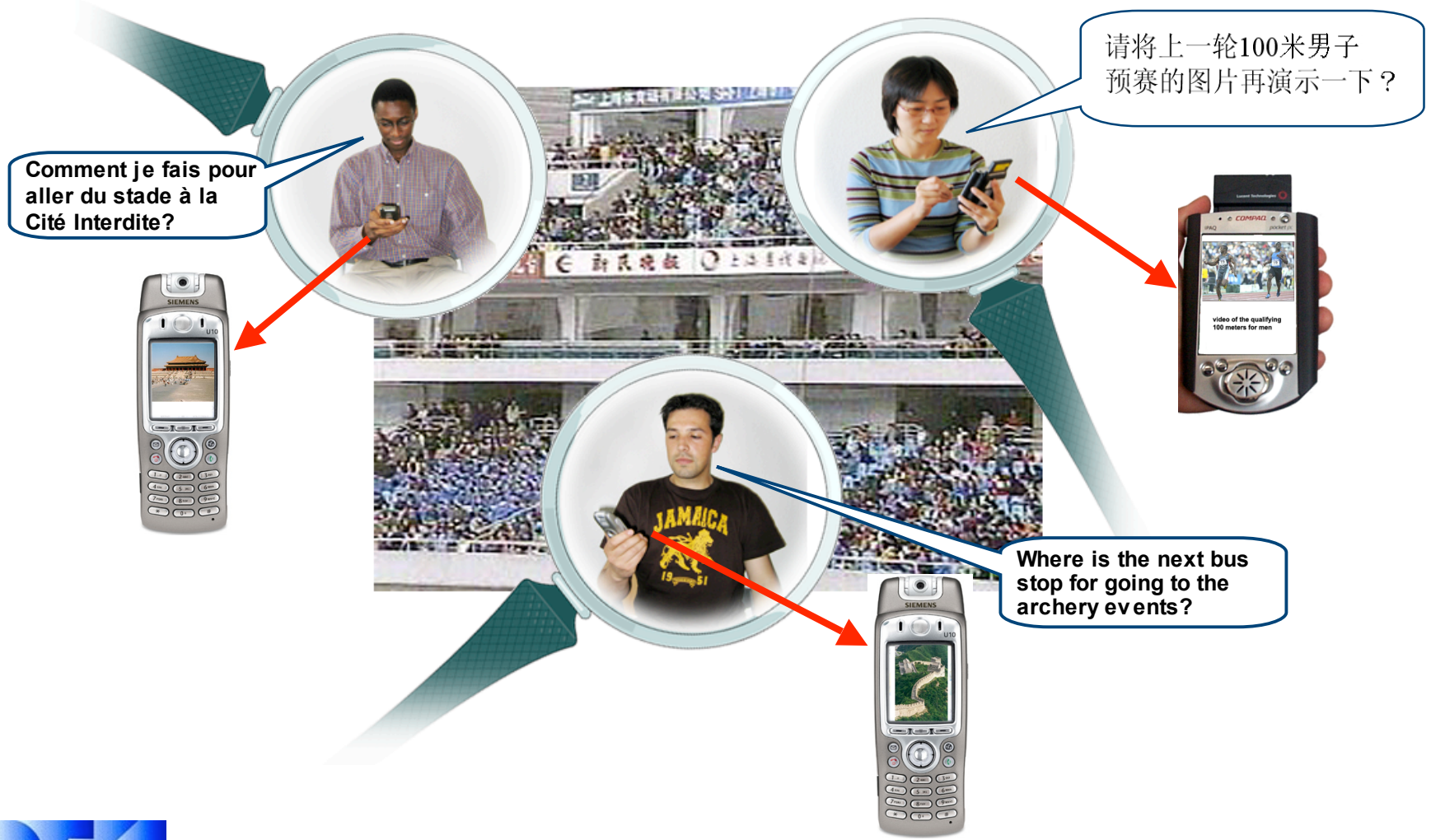
- ☆ A Sino-German Project feeding into Chinese Digital Olympics Projects (viz. Invited Lecture by Weiquan Liu)

- ☆ Four Selected Areas for MT
  - smart dining
  - taxi dialogue
  - emergency assistant
  - general MT web service

A Comprehensive Public Information Services System for the Olympic Games 2008 in Beijing

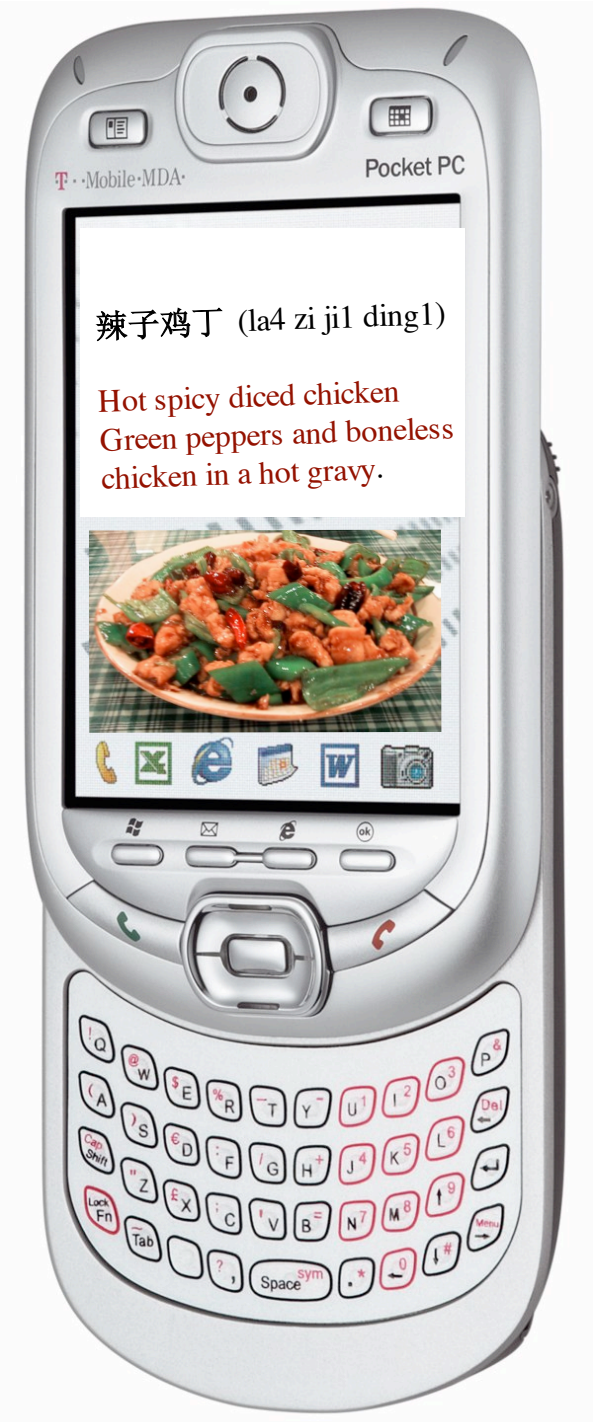
Fraunhofer Institut Software- und Systemtechnik

Multilingual Mobile Multimodal Information Services



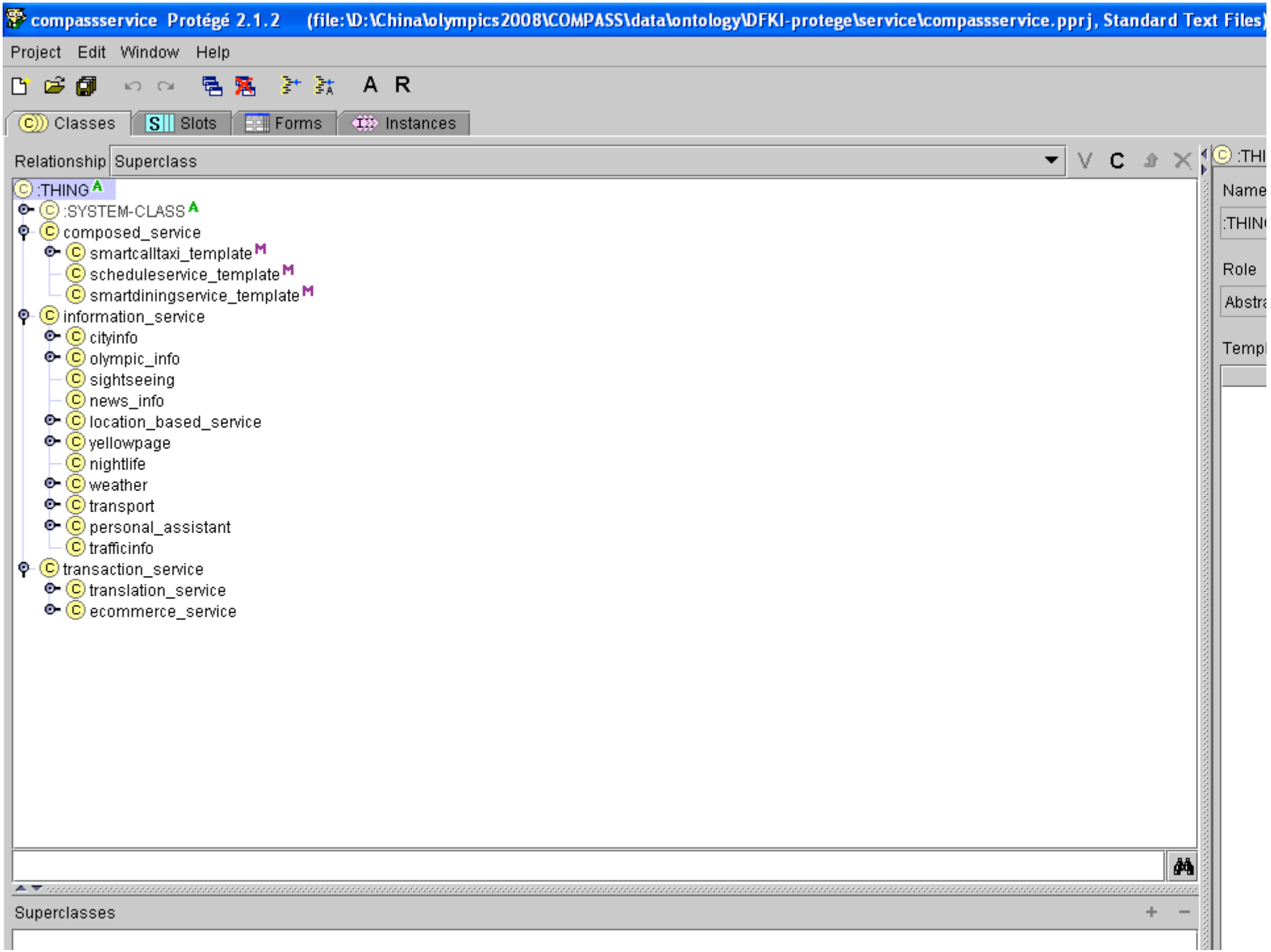
Services

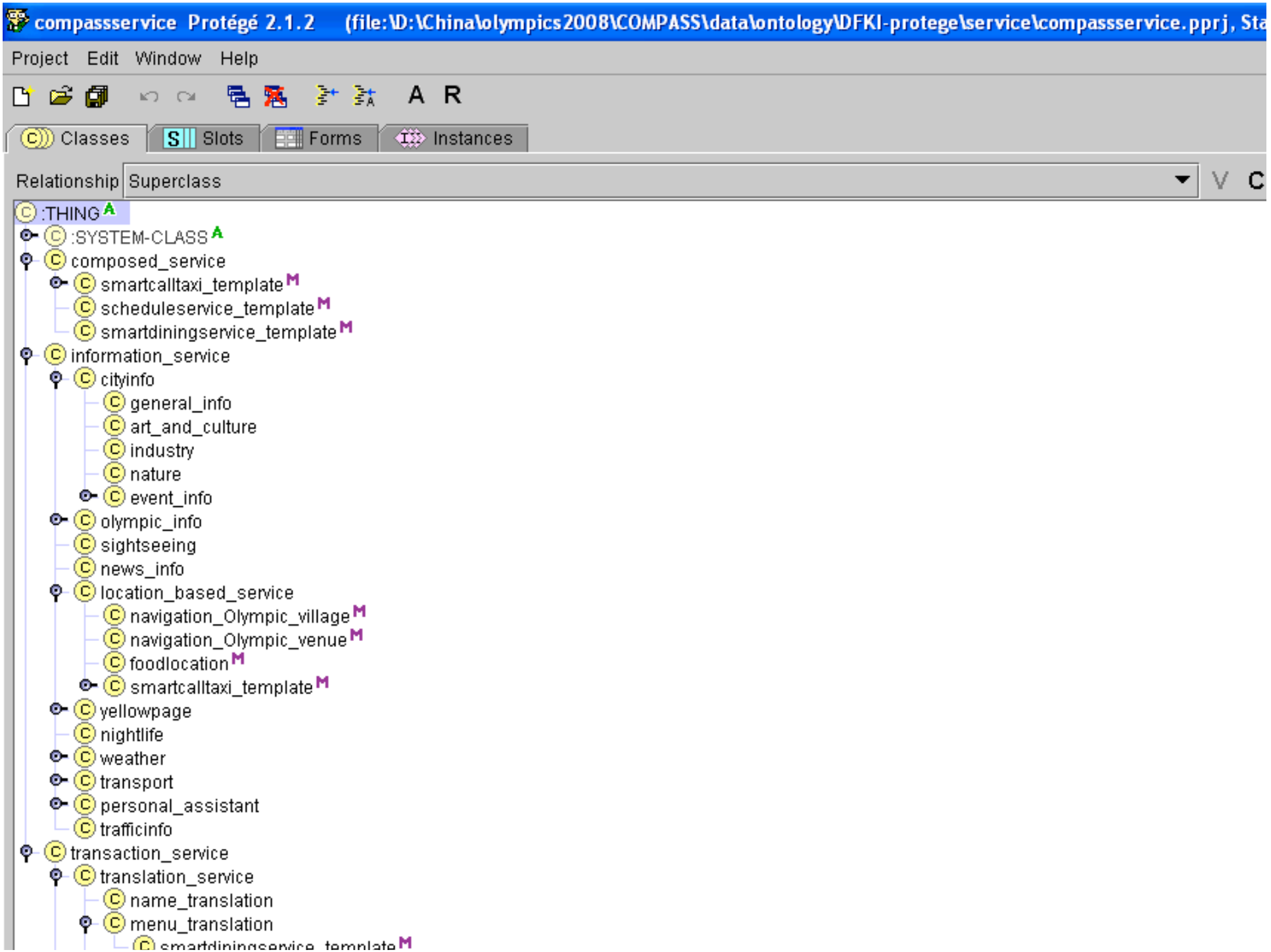
- ☆ translation services
- ☆ city navigation
- ☆ event information
- ☆ crosslingual communication help
- ☆ tourism information
- ☆ smart dining assistance
- ☆ taxi dialogue assistance
- ☆ emergency help
- ☆ olympic news service



Specialized Translations in COMPASS 2008

- ☆ smart dining (Chinese, English, German)
  - help for foreigners
  - food and dining expressions from phrasebook
  - ontology of food items/cuisine
  - instances: dishes from specific restaurant menues
- ☆ taxi dialogue (Chinese, English, German)
  - ontology in preparation (in connection with navigation
- ☆ emergency assistant (many languages)
  - phrasebook-like specialized constrained, fool-proof translation
  - connection to emergency services
- ☆ general MT web service
  - web-based general translation service (HuaJian, LOGOS, Systran and others)


compassservice Protégé 2.1.2 (file:\D:\China\olympics2008\COMPASS\data\ontology\DFKI-protege\service\compassservice.pprj, Standard Text Files)
Project Edit Window Help
Classes Slots Forms Instances
Relationship Superclass
:THING
:SYSTEM-CLASS
composed_service
smartcalltaxi_template
scheduleservice_template
smartdiningservice_template
information_service
cityinfo
olympic_info
sightseeing
news_info
location_based_service
yellowpage
nightlife
weather
transport
personal_assistant
trafficinfo
transaction_service
translation_service
ecommerce_service
Superclasses
Name
Role
Abstra
Temp


compassservice Protégé 2.1.2 (file:\D:\China\olympics2008\COMPASS\data\ontology\DFKI-protege\service\compassservice.pprj, Sta
Project Edit Window Help
Classes Slots Forms Instances
Relationship Superclass
:THING
:SYSTEM-CLASS
composed_service
smartcalltaxi_template
scheduleservice_template
smartdiningservice_template
information_service
cityinfo
general_info
art_and_culture
industry
nature
event_info
olympic_info
sightseeing
news_info
location_based_service
navigation_Olympic_village
navigation_Olympic_venue
foodlocation
smartcalltaxi_template
yellowpage
nightlife
weather
transport
personal_assistant
trafficinfo
transaction_service
translation_service
name_translation
menu_translation

Different Demands for Different Applications

inbound translation for intelligence services, search engines	outbound translation for publishing	outbound translation for information services
indicative/informative translation of web content	publishable translation of documents	precise translation of provided information
broad domain or open domain	narrow to broad domain	narrow to broad domain
robust with inhomoge-neous quality	self-confidence estimation selective with some high-quality portions	maximally robust within area of specialization
usually one translation	output of a few competing realizations can help	one translation

- ☆ parallel processing lines (selection by voting, self-confidence rating, statistical realization ranking)
- ☆ shallow preprocessing and rule-based approaches (POS Tagging, term identification, NED, relation detection)
- ☆ rule-based approaches and statistical postprocessing (realization ranking)
- ☆ disambiguation calls to thesauri, ontologies, inference engines

- lack of funding for large-scale efforts (at least in Europe)

circulus vitiosus

- gap between research systems and top commercial products in rule-based MT

- missing breadth of expertise at most research sites

- ☆ efficient technology evolution requires some continuous copying, spreading and sharing of base-line technology
- ☆ the open-source approach offers such a broad basis for technology evolution
- ☆ statistical base technology for statistical MT is available

- ☆ The open source Giza++ toolkit (Och and Ney 2003) provides the technology for estimating the parameters of statistical translation models from a parallel corpus. Widely used.

- ☆ The Pharaoh decoder (Koehn 2004b) contains technology for producing translations of new texts on the basis of statistical translation models. Pharaoh is provided free by ISI USC but is closed source.

A joined project between

GLOBALWARE

and

German Research Center for Artificial Intelligence GmbH

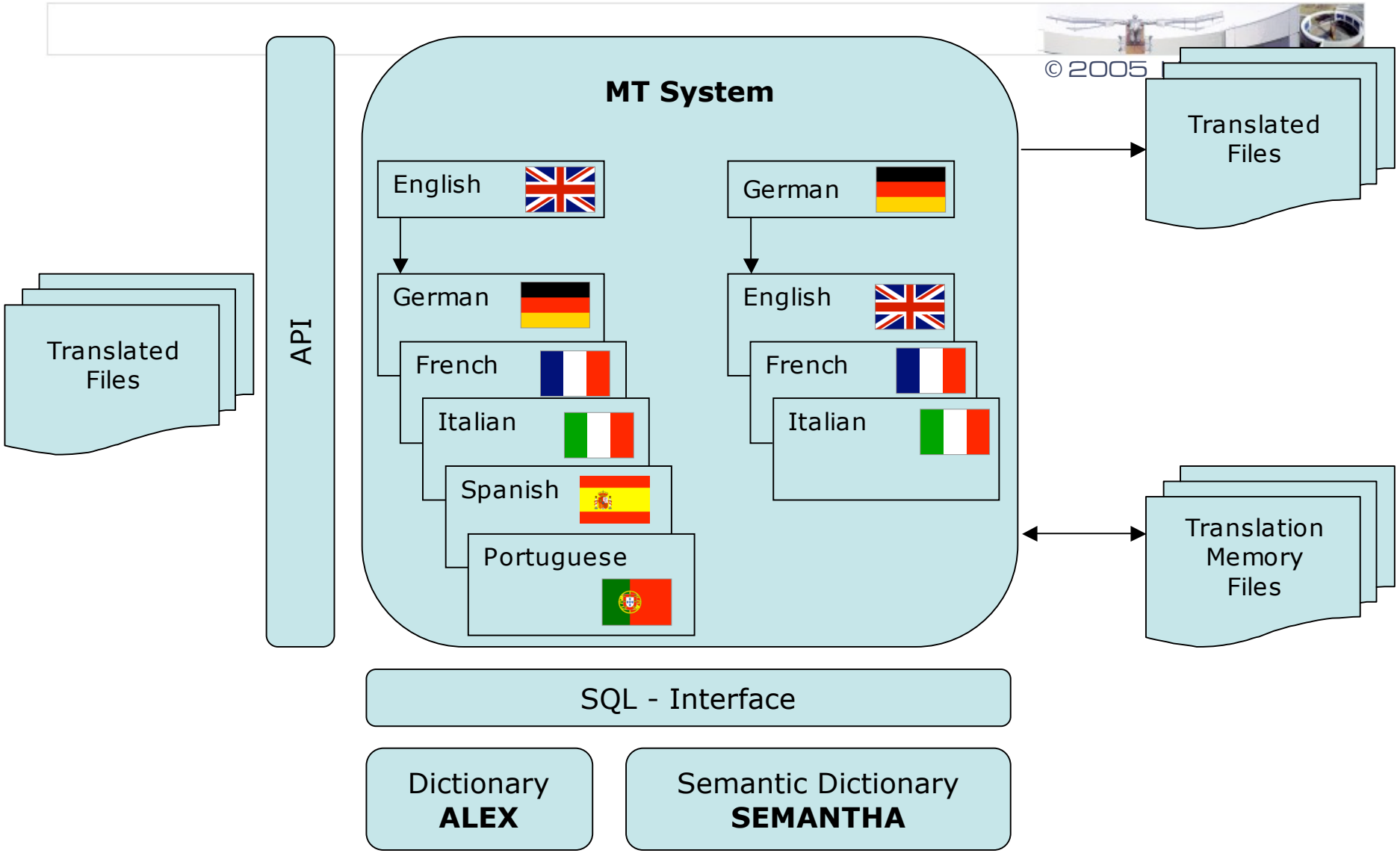
OPEN SOURCE

LOGOS MT

www.logos-mt.com

LOGOS MT

- ☆ Besides Systran, IBM Personal Translator, METAL-Comprendium etc. one of the oldest, largest and commercially most succesful MT Systems
- ☆ 40 year history: Bernhard E. Scott
  started work in 1965 at Computer Technology Inc.
- ☆ LOGOS Corporation was founded in 1969
- ☆ More than 140 Mio Euro development costs


© 2005
MT System
English
German
French
Italian
Spanish
Portuguese
German
English
French
Italian
API
Translated Files
Translated Files
Translation Memory Files
SQL - Interface
Dictionary
ALEX
Semantic Dictionary
SEMANTHA

SCHEDULE

- ☆ September 2005: First complete system without proprietary components finished – announcement at MT Summit X
- ☆ October 2005: Windows and Linux Systems fully functional
- ☆ January 2006: Logos Development Kit will be distributed
- ☆ April 2006: v1.0 will be become available

- ☆ After a long phase of depression, we seem to enter a new exciting phase of progress in MT
- ☆ For the first time, there will be methodologies and accessible state-of-the-art baseline technology for experimentation with combinations of approaches
- ☆ Since we will most likely not find the optimal cognitive architecture in the near future, hybrid approaches, including the clever combination of specialist systems
  will help to understand the problem and build applications

Thank you for your attention...