Language and Communication Technologies: Education & Research at FUB

Raffaella Bernardi

Free University of Bozen-Bolzano

Contents

1. What are LCT?

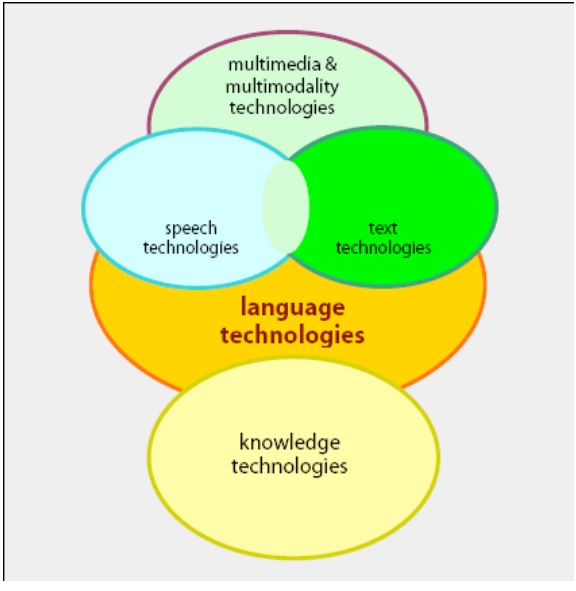
LCT are information technologies specialized to deal with the most complex information medium:

Natural Language

It involves:

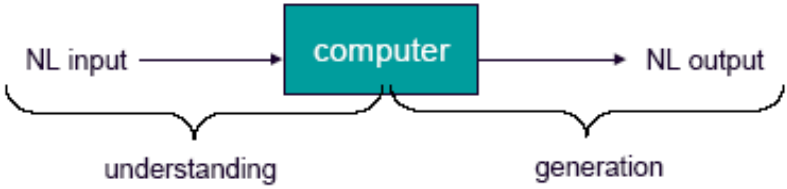
- Text
- Speech
- Knowledge
- Gesture, Facial Expressions
- etc.

1.1. In an image



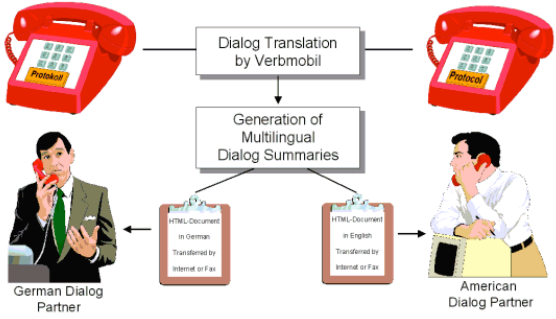
1.2. Goals of LCT

- **Ultimate goal**: To build computer systems that perform as well at using natural language as humans do.
- **Immediate goal** To build computer systems that can process text and speech more intelligently.



1.3. Applications: an example

Speech Recognition and Cross-Language Technologies help communication between speakers of different languages



2. LCT within the EM in CL

Module on LCT Possible courses:

- Computational Linguistics
- Cross Language Information Technologies
- Digital Libraries
- Human Computer Interaction
- Introduction to Linguistics
- Text Processing
- ...

Courses thought by FUB lecturers, researchers from research centers ITC-irst and EURAC, and by companies (CELI, Torino).

Project & Thesis Students can join on-going projects to gain hands-on experience.

2.1. Other activities

EM in LCT The LCT module is part of the **European Masters Program in LCT**: Charles University in Prague, Copenhagen Business School, Groningen University, Roskilde University, Saarland University, University of Amsterdam, University Henri Poincaré, University of Nancy 2, University of Malta, Utrecht University.

http://www.inf.unibz.it/mcs/lct/

LCT Colloquia weekly seminar on LCT. Speakers from FUB, ITC-irst, Trento University/CiMeC, EURAC, and international renown invited speakers. This year seminar: http://www.inf.unibz.it/mcs/lct/seminars-2007.php

LCT Reading Groups overview talks on different aspects of LCT (last years), or more tidily related to students projects (e.g. this year on IQA).

2.2. Examples of Students Projects/Theses

- Luciana Benotti: "Enhancing a Dialogue System through Dynamic Planning"
- Marija Slavkovik: "Constraint Relaxation for IQA"
- Pasquale Imbemba: "A splitter for German Compound words"

3. Research on LCT at FUB

We are working on **Natural Language Interface to Information Systems**.

The final aim is tackled from different perspectives and its subdivided into several projects that hopefully will gather at the end in a unique system.

An example:

Topic Controlled Natural Language for querying, specifying an Ontology.

People: Camilo Thorne (PhD project), Raffaella Bernardi, Diego Calvanese.

3.1. Controlled Natural Language

Problem Natural language access to DB, Ontology (specify, query, update etc..)

Approach Use a suitable fragment of natural language (a **controlled natural language**) [Sowa 2004].

Systems have been proposed that:

- guide the user to formulate his/her question via an ontology that incrementally shows the possible concepts on which the remaining part of the question could be about [Dongilli et al. 2004]
- guide the user via an incremental parser [Bernstein 2005, Schwitter 2004].

Both approaches aim to allow the user to build only those **questions that the system can handle**.

Our proposal Try to answer the question of **which** should be the natural language fragment to be used for such a purpose, and **how** we can define it.

3.2. Natural Language Fragments

Ian Pratt is investigating the semantic complexity of fragments of natural language, i.e. the computational complexity of deciding whether any given set of sentences in that fragment represents a logically possible situation. For instance, given the following words

Verbs	is a	is not a	
Determiners	some	every	no
Nouns	man	...	
Proper Names	Socrates	...	

we can built sentences of the structure below:

Every man is a mortal
Socrates is a man

from which we infer "Socrates is a mortal" that is still a structure built out the lexicon above.

The fragment of sentences built out of this lexicon is called COP.

3.3. Complexity of NL fragments

The FOL meaning representation of the entailment above is:

$$\{\forall x(man(x) \rightarrow mortal(x)), man(socrates)\} \models mortal(socrates)$$

Pratt has proved that COP is PTIME

Fragment	Decision class for satisfiability
COP+TV+DTV	**PTIME**
COP+REL	**NP-Complete**
COP+REL+TV	**EXPTIME-Complete**
COP+REL+TV+DTV	**NEXPTIME-Complete**
COP+REL+TV+RA	**NEXPTIME-Complete**
COP+REL+TV+GA	undecidable

TV	transitive verb, eg. X knows	DTV	transitive verb, eg. X give Y Z
Rel	relative pronoun, eg. who X	GA	general anaphora, e.g. him
RA	restricted anaphora		

3.4. "Which" from the ontology perspective

Which fragment? Our proposal is to merge Pratt's approach with the research mentioned above and use as controlled language for accessing ontologies those fragments with a **desirable computational complexity**.

- Description Logic (DL) are the logics that provide the formal underpinning to ontologies and the Semantic Web.
- DL-Lite is the maximal DL that has the ability to efficiently and effectively manage very large data repositories by relying on industrial-strength relational database management systems (RDBMS). Moreover, DL-Lite can capture the essential features of the most commonly used formalisms for conceptual modeling, such as UML class diagrams and entity-relationship schemas
- Hence, we use a DL-Lite as the starting point to answer the **which** part of our question, viz. to pinpoint the most suitable fragment.

3.5. English lite

The constraints expressed in the TBox are universals. They are of the form $Cl \sqsubseteq Cr$ that translates into FOL as $\forall x.Cl(x) \rightarrow Cr(x)$ and in natural language as

(a) [Every $\underbrace{\texttt{NOUN}}_{Cl}$] $\underbrace{\texttt{VERB\_PHRASE}}_{Cr}$

(b) [[Everyone $\underbrace{[\text{who } \texttt{VERB\_PHRASE}]}_{Cl}$] $\underbrace{\texttt{VERB\_PHRASE}}_{Cr}$]

3.6. English lite: examples

Interesting examples are the ones with relative pronoun (Recall: COP+Rel NP-Complete!):

(1) Everyone who eats left $[\exists Eats \sqsubseteq Left]$

(2) Everyone who knows something left $[\exists Know \sqsubseteq Left]$

(3) Every student who studies left.
$\forall x.(\texttt{student}(x) \wedge \texttt{study}(x)) \rightarrow \texttt{left}(x)$ $[Student \sqcap \exists Study \sqsubseteq Left]$

(4) Every student who is a boy left.
$\forall x.(\texttt{student}(x) \wedge \texttt{Boy}(x)) \rightarrow \texttt{left}(x)$ $[Student \sqcap Boy \sqsubseteq Left]$

(5) Every student who eats something left.
$\forall x.(\texttt{student}(x) \wedge \exists y.\texttt{eats}(x, y)) \rightarrow \texttt{left}(x)$ $[Student \sqcap \exists Eats \sqsubseteq Left]$

(6) Everyone who drinks something and eats something left.
$\forall x.(\exists y.\texttt{drink}(x, y) \wedge \exists z.\texttt{eats}(x, z)) \rightarrow left(x)$ $[\exists Drinks \sqcap \exists Eats \sqsubseteq Left]$

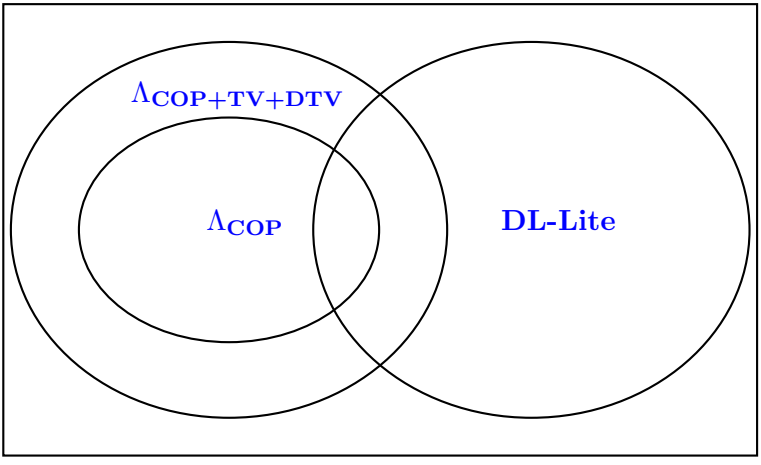
3.7. Relative clauses outside English Lite:

The meaning representations of the sentences below are not in DL-Lite, hence these sentences are outside English Lite.

(7) Everyone who does not know something left $[\neg\exists Know \sqsubseteq left]$

(8) Everyone who is not a boy left. $[\neg Boy \sqsubseteq left]$

3.8. Pratt's NL fragments vs. English lite



3.9. How: Formal Grammar

We have built a grammar able to recognize **only** those sentences whose meaning representation is in DL-lite.

Questions

- Can we be satisfied?
- Can we do more, and define a grammar that recognizes "all and only" linguistic structures whose meaning representation is in DL-lite?
- But how can we define the "all"?
- Would an user be happy in using a Controlled Natural Language?
- How far is this CNL from the sentences that a user would naturally use to access Information Systems?
- Would we ever be able to bridge this gap?

3.10. Other Perspectives

We are:

- **generating** natural language structures from conjunctive query language.

  People: Paolo Dongilli (PhD project), Enrico Franconi, Sergio Tessaris.

- studying re-writing rules from "free text" into some CNL and from CNL (the generated sentence) into more "natural" sentences. Hence, we are exploring **paraphrases** in natural language.

  People: Francesca Bonin (MSc thesis), Raffaella Bernardi, Bonnie Webber.

- developing a logic and statistic based **incremental parses** using Grammar Induction method, that could help guiding the user.

  People: Andrea Bolognesi (PhD project), Raffaella Bernardi, Michael Moortgat.

3.11. Other approaches

We are using statistical based methods to

- develop an **Interactive Question Answering** (IQA) system (BoB http://alice.inf.unibz.it:8080/opac/)

  People: Manuel Kirschner (PhD project), Raffaella Bernardi, Massimo Poesio.

- enhance OPAC with a **multilingual access system** to Library Catalogues (MuSiL: http://pro.unibz.it/opacDocDigger/)

  People: FUB/CS (Raffaella Bernardi, Diego Calvanese, Barbara Plank) CELI (Luca Dini, Paolo Curtoni, Vittorio Di Tomaso), FUB/Library (Elisabeth Frasnelli, Ulrike Kugler.

4. Conclusions

Further info: visit the home pages of the people involved in the LCT-research

- Raffaella Bernardi http://www.inf.unibz.it/~bernardi/
- Andrea Bolognesi (Siena University) http://www.andrea.bolognesi.name/
- Diego Calvanese http://www.inf.unibz.it/~calvanese/
- Paolo Dongilli http://www.inf.unibz.it/~dongilli/
- Enrico Franconi http://www.inf.unibz.it/~franconi/
- Manuel Kirschner http://web.inf.unibz.it/~mkirschner/
- Barbara Plank: http://www.inf.unibz.it/~plank/
- Massimo Poesio (Trento University) http://cswww.essex.ac.uk/poesio/
- Sergio Tessaris http://www.inf.unibz.it/~tessaris/
- Camilo Thorne http://www.inf.unibz.it/~thorne/

4.1. MSc students and Research Centers

- Francesca Bonin
- Pasquale Imbemba http://www.geocities.com/imbemba/
- Marijn Schraagen
- Marija Slavkovik
- ITC-irst: http://tcc.itc.it/
- EURAC: http://www.eurac.edu
- CELI: http://www.celi.it/