(Towards) A Roadmap for Controlled Translation

Steven Krauwer
Utrecht University / ELSNET
steven.krauwer@elsnet.org

Overview

- About ELSNET
- About The ELSNET Roadmap
- What ELSNET has to say about the MT Roadmap
- What I have to say about MT and CT
- What you have to say about the CT Roadmap
- Scenarios for the future
- Questions
- Concluding remarks

About ELSNET

- European Network in Human Language Technologies (http://www.elsnet.org)
- Created in 1991; now some 140 member organisations in 29 countries
- Fully funded by the European Commission
- Hosted by Utrecht University
- Objectives:
  - Bring together language and speech communities
  - Bring together academia and industry

Membership

- Open to organisations (public and private) active in R&D in language or speech technology
- Conditions on the site
- Membership is free
- What you get:
  - Influence on actions
  - Support for activities (no grants)
  - Quarterly newsletter
  - Discounts
- What you give: active participation

What we do

- Training (e.g. schools, workshops, courses, curriculum development)
- Information dissemination (e.g. website, paper newsletter ELSNews, mailing lists)
- Resources and evaluation (e.g. workshops)
- Roadmapping (e.g. website, documents, workshops)

About the Roadmap

- Requested by the EC
- Responsible: DFKI (with Pisa and Utrecht)
- Broadly supported view of our future
- Identifies major challenges and intermediate milestones
- Used by researchers and funders to identify common priorities in order to focus efforts and exploit synergies
- A roadmap is not a prediction but rather an instrument to make the future happen sooner

Some problems

- Language and speech technology is very diverse, as opposed to e.g. microprocessors (faster, smaller, less power consumption) or high temperature electronics (linear scale)
- Field is very dynamic
- External developments may have a huge impact (mobile applications)
- Availability of language resources major bottleneck (slow and expensive, many languages)
- Consensus necessary, but hard to create

Status

- Should cover in principle all sub-fields of language and speech technology
- Overview of what we have on http://www.elsnet.org/roadmap.html
- 7 workshops; 2 documents
- Formal approach (object oriented)
- Graphical representation can be found at http://elsnet.dfki.de (far from complete)
- Community invited to comment and contribute

MT: Some milestones

- 2003: task oriented interpretation
- 2004: portable MT systems
- 2005: spoken sentence-based translation
- 2007: usable ontologies for many domains
- 2007: spoken language MT systems
- 2008: controlled language MT systems
- 2008: translator's workbench
- 2010: speech/text translation

MT: Research problems

- Developing a formal theory of translation
- Developing a semantic theory
- Eliminating the knowledge acquisition bottleneck
- Using translation memories (bi-texts) and machine translation together in a product
- Creating permanent shared language repositories (sharing), including huge, word aligned multi-texts
- Moving towards a theory of cross-lingual communication aids for situation dependent solutions

MT: User dreams

- Language plug-ins for mobile phones (for transactions rather than full fledged interpretation)
- Help with the hard part of foreign languages.
- Large MT evaluation from user perspective.
- Standard control menu language (for cross-language communication by means of small menu driven devices)
- Cross-lingual sign-reading eyeglasses (foreign language signs or messages are read by a small camera, and the translation is projected in the user's glasses)
- Learning from user feedback (via post-edition tools), and predicting user needs, constructing user models
- Web search and translation with CLIR.
- Automatic stenography (TV, conferences)

MT: Industry challenges

- Language plug in for cell-phone, but as a service
- Ways to stick language learning books into MT systems
- Using TM (bi-texts) & MT together in a product
- Coverage of Minority languages.
- Massively annotated multi-text.
- Exploiting markup.

Where are we now?

- Incomplete and inconsistent calendar of milestones
- Unstructured wish lists from the researcher, user and developer/provider perspective
- Nothing sensible on controlled languages

Where I stand

- I'm in MT, not CL or CT
- Do I believe in full-fledged MT?
- Strong believer in divide and rule
- Strong believer in CL
  - Engineers are lousy writers
  - Writers are lousy engineers
- Many objections to CL emotional

What you say about CT

- My analysis of this conference (23 papers)
  - 9 very interesting but unconnected to CL
  - 4 about use, construction, checking, editing
  - 4 addressing philosophical questions
  - 3 special (dialogue systems, inferencing, knowledge acquisition)
  - 2 where the system is taking control
  - 1 data assisted CT

Philosophical questions

Neat work, messy resulting picture:

- Relation between CL for humans and for MT:
  - Ursula Reuther (translatable implies readable)
  - Marina Vassiliou (readability improves translatability; why not tune CL to the MT system)
  - Margrethe Moeller (humans/experts/MT)
- Universals for CL (for humans and MT):
  - Sharon O'Brien (no evidence other than short sentences)

Specials

Not really CT, but open new perspectives:

- Let system manipulate user in dialogue
  - Arendse Bernth
- CL for knowledge representation and reasoning
  - Jana Sukkarieh
  - Rolf Schwitter

The system takes over

- Symbolic input to multilingual CL
  - Richard Power
- Symbolic input to spoken output in one language
  - Cristina Vertan

Data assisted CT

- Exploitation of translator knowledge in CL:
  - Michael Carl
- Statistical and example based methods seem to dominate most other CL conferences, but not here. Why?

Reactive vs. proactive

- Most editors and checkers seem to be reactive
- Proactive examples:
  - Arendse Bernth (dialogue control)
  - Richard Power (options)
  - Rolf Schwitter (look-ahead)
  - Cristina Vertan (menus)
- Is this a new/good trend in CL?

Some scenarios

- Scenario A:
  - Stand-alone MT remains at present level
  - Usable MT continues to be surrounded by CL and post-editing; CL tuned to MT system
  - CL for for humans further improved, based on better understanding of universals

Some scenarios

- Scenario B
  - Stand-alone MT remains at present level
  - CL for MT and for humans can find common theoretical basis
  - Usable MT continues to be surrounded by CL and post-editing, but CL more generic

Some scenarios

- Scenario C:
  - Stand-alone MT remains at present level
  - CL for humans AND usable MT based on multilingual generation under full control of system

Some scenarios

- Scenario D:
  - Stand-alone MT improves dramatically
  - CL for MT evaporates, input control on level of spelling/grammar/format is handled by next generation word processor
  - Some post-editing may be necessary for export translation
  - CL for humans continues to evolve or is completely taken over by multilingual CL generation on the basis of symbolic input

Questions

- The example scenarios are not exhaustive, but where do we expect/hope/fear that we are heading?
- What to focus on and where to invest?
- What are the requirements in terms of language resources?
- Is stand-alone MT a real factor (that could make CL superfluous), or can we just ignore it?
- Do we have the appropriate (and affordable) evaluation methodologies?

Concluding remarks

- The ELSNET roadmap for CT and MT is still embryonic
- Scenarios may help structuring the universe
- Important parameters for CT are:
  - What is going to happen to MT
  - Can we get a better theoretical understanding of CL for humans (and systems)
  - How strong is multilingual generation going to be

Thank you