A Multilingual Europe -- Can we Really Handle This?

Steven Krauwer
Utrecht Institute of Linguistics / ELSNET
steven.krauwer@elsnet.org

Purpose

Purpose of this presentation:

- not to answer the question because it is a ‘philosophical question’
- not to tell you what I have done
- but to discuss where the field is or should be moving
- based on the ELSNET roadmap activities and my own opinions and beliefs

Overview

- The ELSNET roadmap
- Setting the problem scene
- Current MT
- Stating the problem
- Divide and rule
- Concluding remarks

What is ELSNET

- European Network in Human Language Technologies (ca 140 academic and industrial member organisations)
- Funded by the European Commission
- Created in 1991
- Objectives
  - bringing together the language and speech communities
  - bringing together academia and industry
  - facilitating R&D in language and speech technology
- Info: elsnet@elsnet.org http://www.elsnet.org

What is a roadmap

- A broadly supported vision of where our field is going (research, technology, market)
- which should help us (= researchers, developers, providers, funders, educators) to
  - identify main challenges
  - set intermediate milestones
  - concentrate efforts
  - measure progress and (if necessary) adjust goals

Some words of caution

- It reflects expectations rather that predictions or commitments
- It is highly dynamic in that changes in funding priorities may have a severe negative or positive impact
- It is highly dynamic in that external factors (evolving technologies and markets, political crises) may lead to dramatic changes

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)
- Research perspective still overrepresented
- Community invited to comment and contribute
- MT still in information gathering phase

MT: Some milestones (anno 2000)

- 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 paid 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.

Status of the MT Roadmap

- Incomplete and inconsistent calendar of milestones
- Unstructured wish lists from the researcher, user and developer/provider perspective
- Still gathering information and trying to structure it

Setting the scene (1)

All EU citizens will be living in one

- economic space
- cultural space
- monetary space (eventually)
- information space
- political space (within certain limits)
- touristic space
- entertainment space
- …

Language barriers constitute a major obstacle; how bad is it?

Setting the scene (2)

- EU has
  - 15 member states, with 11 official languages (plus quite a few ‘unofficial languages’)
  - 10 new member states with (at least) 10 new official languages
  - 3 applicant countries with at least 3 extra languages
- Europe has
  - 17 other countries with ??? other languages

Setting the scene (3)

The Ethnologue (http://www.ethnologue.org):

- Europe: 230 languages
- The Americas: 1013 languages
- The Pacific: 1311 languages
- Africa: 2058 languages
- Asia: 2197 languages

Language barriers are real obstacles

- It is not true that most people speak English
- It is not true that most people have easy access to language learning facilities
- Human translators are adequate, slow and expensive
- Human interpreters are adequate, fast and terribly expensive
- Human intervention is not always practical
- Can MT offer a solution?

Current MT

- State of the art MT is after 50 years as good (or as bad) as what you get when you use the AltaVista or Google translation facility:
  - poor, at times incomprehensible or even ridiculous,
  - but always better than nothing if you don't know the language at all
- No real step forward for the last 10-15 years

MT (spoken or written) is hard

For traditional (rule-based) approaches:

- No language is completely described
- No language has a complete dictionary
- We have no systematic knowledge of correspondences and discrepancies between languages
- Ambiguity resolution is crucially dependent on real world knowledge, which is hard to accommodate
- Adding speech makes the nightmare worse

MT (spoken or written) is hard

For statistical methods:

- We need vast amounts of parallel bilingual corpora
- We need massive storage and computation power (the only problem we can be sure will go away if we are patient enough)
- Adding speech makes the nightmare worse

But: they made a quick start and if it works they allow for quick deployment and don't need world knowledge

Stating the problem

- Option 1: The problem is that our MT systems aren't good enough. Let's invest more in MT system R&D, both statistical and rule based until we know how to do it
- Pro:
  - great if and when we get there …
- Con:
  - … but what if we don't?
  - and what do we do in the meantime?

Stating the problem

- Option 2: The real problem is that language barriers are in the way for successful communication, where each communication situation may be different
- Pro:
  - allows for ‘divide and rule’ approach
  - allows for different success criteria
- Con:
  - instead of one big problem you have to solve lots of smaller problems

Partitioning the problem

- Don’t try to imitate the human translator
- Perfect translation is not a necessary condition for successful communication
- Observe that cross-lingual communication situations are different, and not symmetric
- Don’t try to find one solution for all situations, but rather for each typical situation a suitable solution
- Be pragmatic: don’t be dogmatic about methods

Cooperativity matrix

- 2 parties:
  - *The customer*: the party who needs something (information, hotel room, potatoes)
  - *The provider*: the party who has it
- 2 attitudes:
  - Cooperative (is prepared to make an extra effort to make the communication successful)
  - Uncooperative (is not prepared or able to make the extra effort)

The matrix

Cooperative customer C Cooperative provider P	Cooperative customer C Uncooperative provider P
Uncooperative customer C Cooperative provider P	Uncooperative customer C Uncooperative provider P

Coop C⇔Coop P (1)

Examples:

- Conferences (speaker and listener use foreign language).
  Technology: language learning facilities
- Mediated speech translation (cf Verbmobil)
  T: restricted domain MT, common language to solve problems
- Chatting
  T: poor MT (iteration to solve problems)

Coop C⇔ Coop P (2)

Priorities

- Go for language learning facilities
- And require every EU citizen to speak at least two other EU languages

Coop C⇔ Uncoop P (1)

Examples

- Tourism: C uses local language
  T: language learning facilities
- Tourism, military: C uses hand-held devices
  T: electronic dictionaries or phrase books, possibly with speech in- and output
- Web tourism (at most gisting)
  T: state of the art PC translation

Coop C⇔ Uncoop P (2)

Priorities:

- Intelligent phrasebooks (with speech IO and translation capabilities)

Uncoop C ⇔ Coop P (1)

Examples:

- Information for hotel guests
  T: Multilingual generation from e.g. tabular information or conceptual representations
- Business correspondence
  T: Foreign language authoring tools
- User manuals
  T: Controlled languages
- Touristic transactions
  T: Foreign language dialogue systems

Uncoop C ⇔ Coop P (2)

Priorities (1)

- Controlled languages (design and authoring tools):
  - Should guarantee translatability
  - Should guarantee intra company consistency
  - Should guarantee decently structured text

Uncoop C ⇔ Coop P (3)

Priorities (2):

- Quick deployment spoken dialogue systems
- This requires
  - Good ergonomic design (minimize user irritation)
  - Widely used conventions (improve user behaviour)

Uncoop C⇔ Unc P (1)

Examples:

- Sensitive (e.g. political) discussions
  T: none, just human interpretation, possibly with technological support
- High quality translation needed
  T: human translation with productivity tools (translation memories);
  or (poor) MT with post-editing (VERY cost effective!)

Uncoop C⇔ Unc P (2)

Priorities:

- Further development of MT systems
- Look at hybrid methods
- Better integration of MT (with post-editing in workflow)
- Fast customization methods

Concluding remarks

My main beliefs:

- Divide and rule
- Give high priority to
  - Language learning
  - Controlled languages
  - Spoken dialogue systems
- And don’t give up on ordinary MT