Building Term Dictionaries for Machine Translation in Practice: A User Experience

Annelise Bech

Abstract

The paper starts with a brief outline of the machine translation set-up at Lingtech and the work-flow in the organisation. In-house the main task concerns preparing the patent texts for machine translation; the bulk of this work actually consists of identifying and coding technical terms and expressions. To this end we have two computational linguists, who are also in charge of creating and maintaining term databases for the various domains the texts concern. We intend to focus on our experiences and discuss problems and solutions from a practical point of view.

Annelise Bech

Ms Bech has been involved in language technology for more than ten years. She worked on the Eurotra project from 1986, partly as a member of the Danish language group and partly as a member of a designated specialist group with the task of designing and implementing the Eurotra formalism. Ms Bech also worked with knowledge representation and text understanding during her stay as research fellow at SRI International in California.

Since the creation of the Danish Centre for Language Technology, (CST), she has worked there as project manager of various language technological projects, most notably the development of the PaTrans machine translation system. Ms Bech has worked as a project manager of system development projects at Ramboll; and in November 1995 she took up an appointment as Director of Lingtech A/S.

Lingtech A/S

Lingtech A/S is a specialised translation company established by the patent agency Hofman-Bang & Boutard, Lehmann & Ree A/S. Lingtech has specialised in the translation of patent texts from English, German and French into Danish, and the company translates a total of about 8 million words per year. Lingtech has a core staff of 8 linguists and approximately 50 highly qualified technical experts working on a freelance basis. As a translation company, however, Lingtech is quite exceptional and a pioneer, in that since 1994 the PaTrans machine translation system has been used for English to Danish translation. PaTrans, which exploits Eurotra technology, was specifically designed and developed for Lingtech by CST. Ever since the system was applied in production, Lingtech has seen a steady increase in the number of words translated by the system every month and a level of reduction of more than 50 percent in the translation cost per word when compared to the costs for manual translation. Currently, it is expected that this year will see some 70 percent of the suitable English patent texts being machine translated. These results have led to the planning of further extensions of Lingtech’s application of machine translation.

Annelise Bech, Lingtech A/S
Vesterbrogade 24, DK-1620 København V., Danmark

Tel. +45 33 25 71 71, Fax. +45 33 25 61 71
E-mail: lingtech@login.dknet.dk

(Outline of the Presentation)

- Lingtech and the MT scenario

- PaTrans
- Work-flow

- Term dictionaries

- Definition
- Organisation

- Building term dictionaries

- Strategy
- Problems

- Conclusion

Lingtech:

- Specialised in the translation of patent texts
  English, German and French into Danish

- Bulk of English to Danish translation by MT
  - 70 - 75 % of texts (2 - 2.5 mio. words)
  - cost savings of more than 50 %

The PaTrans MT system

- purpose-designed system for Lingtech
- developed by CST exploiting Eurotra basis
  - translation kernel (Pok)
  - editor (PaEd)
  - term coding tool (PaTerm)

- used in production at Lingtech since 1994

The work-flow

- Registration and OCR scanning of text

- The pre-editing phase:
  Format conversion (WP to PaEd)
  Marking-up
  Dictionary look-up
  Alphabetised check-list
  Simple and multi-word terms

- Term coding

- Batch translation of text

- Post-editing and language revision

Dictionaries

- The general dictionary
  - not maintained by Lingtech

- Term dictionaries
  - number and contents defined and maintained by Lingtech
  - the PaTerm coding tool; nouns, verbs and adjectives

- What is a term?
  1. A single or multi-word expression with a specific translation within a domain
  2. Any single or multi-word expression which we need to add to the lexicon

     component
     abrasion resistant
     fox-shaped

Organisation of term dictionaries

- Patent texts classified according to subject field codes

- The original idea:
  - many specialised dictionaries
  - very fine granularity

- In practice:
  - two 'main' term dictionaries: chem. and mech.
  - three 'supplementary' ones

- Why?
  - linguistically very hard to define
  - various system constraints
  - priority and interaction

    nut
    locking_nut
    fox

- ... 'then you only need to extend the dictionaries. No big deal!'
  In practice it is rather a big deal, though
  Why?
  - real life - size and coverage imperative
  - no off-the-shelf solution to be bought
  - a costly and time consuming task

- terminological homo- or heterogeneity
- the over-all cost-efficiency of MT
- the ‘future value’ of coded terms

- The applied strategy
    - on a text-by-text basis
    - identification and coding of new terms
    - critical selection of texts
    - hitting the right balance between coding and throughput of words

- Some reasons why it is a big deal
    - Coding a term is pretty fast, it’s the finding which is hard

- The two main tasks:
    - identifying new source terms
    - target language translation

- The checklist and single word units
    - categorial ambiguity

      space noun/verb

- The checklist and multi-word units

  ball
joint
....
ball_joint

  front_elevational_view
side_elevational_schematic_view

Interactive concordance facility; but one step

- Finding the translational equivalent
    - quality and consistency
    - expert knowledge; reference material
    - standardisation and corrections

A guiding principle

Work smarter, not harder!

Everyday pragmatics

- Keywords: structured extension of lexicon
    - maximise throughput
    - cost-efficiency

- prefer texts from well-covered fields

- prefer longer to shorter texts
  - from 2.5 to 1.2 - 3.5 to 1.3 percent [ratio *new terms*/*total words of text* for chem. and mech. respectively]

- for new subject-fields prefer short texts

Conclusion

- common format resources ?! In practice ....

- investment, pay-back, added value

- relative terminological exhaustivity
  - figures ?

- needs and market
  - linguistic strategies
  - quantitative measures

- need for invention and integration of tools
  - higher degree of automation
  - proactive/predictive tools
  - reactive/evaluative tools