Translation Technology: some uses of alignment

Cambridge Conversations in Translation
6 February 2017

Andrew Rothwell, Swansea University
Translation Theory abhors alignment

- Source Text priority:
  - Roman Jakobson, ‘On Linguistic Aspects of Translation’ (1959)
  - Vinay and Darbelnet's translation procedures (1958)

- Shift to target:
  - Nida’s Dynamic equivalence (1964)
  - Skopos and functionalism (Reiss 1971, Vermeer 1987)

- ‘Cultural Turn’ – Bassnett and Lefevere (1990):
  - Postcolonialism (Bhabha, 1994)
  - Foreignization (Venuti, 1995)
  - Gender (Simon 1996)
  - Etc...
Three computational approaches to equivalence

• **A: Rules ('classic' MT)**
  - Taggers, parsers, grammars, lexicons
  - SL analysis, Transfer, TL generation

• **B: String matching (CAT)**
  - Translation memory
  - Termbase
  - Fuzzy matches scored as % by Levenshtein (edit) distance

• **C: Statistics (modern MT)**
  - Example-Based Machine Translation
  - N-grams
  - Word-based, then phrase-based SMT
IBM 701 as used in Georgetown experiment (1954)
‘Classic’ MT: linguistic rules & lexicons - Vauquois pyramid (1968)
CAT import-export processes

- **ST import:**
  - File-type (HTML, Word, Powerpoint, InDesign etc.) recognised and filter applied to *separate translatable text from codes/tags* (stored separately)
  - Text is *segmented* according to user-definable rules (punctuation)

- **TT export:**
  - *Codes/tags re-inserted* to produce a TT with the same layout (styles, columns, tables etc.) as the ST
TM fuzzy matching

- Segment: a **meaningful unit of text**:
  - Sentence of body text
  - Title, heading, table cell etc.

- CAT tool compares new SL segment with TM, and finds **matches** based on:
  - Text, in-line formats (bold, italic, footnotes etc.)

- Differences are highlighted as points for the translator to change

- **100%, 101% Context Match; 75% etc. match...**

- **Fuzzy match threshold**: user-defined, normally around 70%

- Quoting and invoicing implications: different rates for
  - New segment translations (100%)
  - Fuzzy matches (variable)
  - 100% match (review only)
Termbase
(SDL MultiTerm 2015)
• 1970s: First commercial systems, e.g. Météo, Systran, Logos, METAL, Trados

• 1978: **Systran** installed at the European Commission; start of EUROTRA

• 1990s: Start of **Statistical MT (SMT)**

• SMT: **Google Translate launched 2006** – already > 10 years ago

• > **99% of translation worldwide is now by MT** (TAUS, p.74): ‘Everyday, in the year 2016, machines translate more than 250 billion words’ (TAUS, p. 10)

• ‘Translation will be available on every screen, in every app and on every signboard’ (p.11)

• **S2S** (p.27) - Google Voice Translator, Microsoft Skype Translation

• **Fully Automatic Useful Translation (FAUT)** by 2030?
SMT (Statistical Machine Translation) works by:

1. **Training monolingual corpora** to generate language models.
2. **Parallel corpora** for the translation model.

**Parallel Corpus**:
The corporation has been estimated to run more than one million pages in data centers around the world to process over one billion search requests and about twenty-four of user-generated data each day. December 2012, Alexa listed as.

**Monolingual Corpus**:
started functioning in 1929 and established the tradition of large exhibitions and trade fairs held in Brno, and nowadays also ranks among the sights of the city. Brno is also known for hosting big motorbike and other races on the Masaryk Circuit, a tradition established in 1930 in which the Road Racing World Championship Grand Prix is one of the most prestigious races. Another notable cultural tradition is an international fireworks competition.
Links to videos and presentations here

TranslatingEurope Forum 2016
Focus on translation technologies
27 and 28 October 2016
Recent MT developments

- **CASMACAT** (Philipp Koehn, U. of Edinburgh): ‘Beyond Postediting’ (project 2011-14)
  - Interactive translation prediction (next 3 words) – like Autocomplete
  - Adaptive translation models
  - Sentence-level confidence measure (ranking against TM)
  - Incremental updating of MT engine with corrections from translator
  - Word alignments
  - Bilingual concordance
  - Thesaurus

- **MATECAT** (Alessia Ridoni, Translated.net)
  - MyMemory (50m web-crawled segments)
  - Projected tag placement (MT works out where to put them)
  - (future) Automatic PE (learning from user corrections)
  - (future) Automatic project creation
  - (future) Context-aware MT, automatic engine creation for domains
Convergence of CAT and MT

- **CAT applied to MT:**
  - SMT engines are trained on aligned corpora, a.k.a. translation memories (TMs). These can be human-produced or Web-crawled.
  - Humans correct MT output (e.g. Microsoft Translator with Feedback).
  - MT confidence scores allow ranking alongside TM proposals.

- **MT applied to CAT:**
  - Sub-segment matching – word and phrase alignments, getting more leverage from TM, using rules and/or statistics.
  - Autosuggest dictionaries (SDL Trados).
  - Grammar and syntax rules deduced from TMs.

- Martin Kay (1980) suggested that incremental improvements to CAT might eventually lead to effective MT.

- New paradigm: Machine Translation Memory (MTM) (Tomáš Svoboda, 2014).

- Demo of Memsourse Cloud...
Against:

- Lack of repetitions
- No requirement for consistent translations
- Segmentation can restrict stylistic freedom

For:

- Avoids accidental omissions!
- Avoids data loss (it’s always in the TM)
- No need to carry the book around
- Makes revision easier
- Particularly useful for a retranslation...
- Deeper reading of the ST

‘Translation is an exacting practice, at once critical and creative.’
Emile Zola (1840-1902)
Critical evaluations and translations of *La Joie de vivre*

- **French review:** ‘l’un des meilleurs romans d’amour écrits sous la Troisième République’ (survey in *Le Figaro*, cited in Fanzén, 1958: 9)

- **English review:** ‘Mainly a colourless lifeless novel, the story introduces ten-year-old Pauline as an impossibly angelic child who grows into an impossibly angelic adult [... ] This smell of burning martyr – which reminds me of Dickens in his worst sentimental moments – continues throughout the novel.’ (Anon. blog, 14 June 2009)

- Translation alignment gives a close-up view of this disparity


- Jean Stewart (1955) *Zest for Life* – long out of print

- Me (forthcoming), *The Bright Side of Life* (Oxford World’s Classics)
Alignment with LF Aligner
Insights from alignment

- Vizetelly sometimes splits or joins Zola’s sentences - interesting options for the re-translator
- Contemporary solutions to problems of Zola’s style (numerous substantivized abstractions, over-use of pluperfect tense)
- Conventional expectations of late-C19 English novel style:
  - Tendency to verbose amplification
  - Frequent explicitation of spatial relations (e.g. as characters move around the house)
  - Refusal of *style indirect libre* (Goethe, Jane Austen and Flaubert)
- Word-level view of Victorian morality and ideology through Vizetelly’s self-censorship
- Questions:
  - Why don’t I want to translate it that way?
  - What will my readership expect?
Re-translating

La Joie de vivre

in MemoQ
Vizetelly’s unmentionables

- Feminine underwear, undressing, nudity, body parts
- Puberty, menstruation
- Female desire (lust, jealousy)
- Physical intimacy (or even contact), bedrooms
- Feline reproductive behaviour
- Pregnancy and childbirth (‘enceinte’, ‘accoucheuse’, ‘state of health’ – ‘midwife’ never used)
- Illicit sexual relations (marital infidelity, prostitution, child mothers)
- Physiology (Lazare’s medical textbooks):
  - ‘machine humaine’ (2:384), ‘mécanique de la vie’ (3:101), ‘fragilité du mécanisme’ (7:300); Louise’s delivery in Ch. 10
Do you wish to revert to paragraph segmented files, or use the sentence segmented versions?

Segment numbers before and after segmentation:

French: 264 → 730
English: 113 → 323

- The segmenting seems to have gone well, so I’ll use the sentence segmented texts
- Revert to the paragraph segmented versions

Note: you should revert to the paragraph segmented files if the segmentation pushed the files badly out of balance (they had a similar number of segments before but not after), especially if (one of) the files hardly gained any new segments.
Chapter 10: Louise’s labour

- 20 pages of extreme emotion and gynaecological detail in the ST;
- Vizetelly’s TT summary:

It was one of those dread hours when life and death wrestle together, when human science and skill battle to overcome and correct the errors of Nature.

More than once did the Doctor pause, fearing a fatal issue.

The patient's agony was terrible, but at last science triumphed, and a child was born.

It was a boy.
Consequences of EAV’s self-censorship

- Pauline’s character and motivations grossly simplified
- Most of the ‘love triangle’ drama flattened
- Pauline’s agency (narrative, moral) weakened
- Zola’s portrayal of Pauline as a woman is betrayed:
  - Bodily experience
  - Emotions
  - Self-education
- Theme of physiology and its link to Naturalism obliterated
- Victorian morality experienced in minute detail
D’un geste violent, elle fit glisser son jupon, enleva sa chemise ; et, nue maintenant, elle se contemplait encore. Ce n’était donc pas pour elle cette moisson de l’amour ? Jamais sans doute les noces ne viendraient. Son regard descendait de sa gorge, d’une dureté de bouton éclatant de sève, à ses hanches larges, à son ventre où dormait une maternité puissante. Elle était mûre pourtant, elle voyait la vie gonfler ses membres, fleurir aux plis secrets de sa chair en toison noire, elle respirait son odeur de femme, comme un bouquet épanoui dans l’attente de la fécondation. [...] Ah ! misère ! la pluie rouge de la puberté tombait là, aujourd’hui, pareille aux larmes vaines que sa virginité pleurait en elle. Désormais, chaque mois ramènerait ce jaillissement de grappe mûre, écrasée aux vendanges, et jamais elle ne serait femme, et elle vieillirait dans la stérilité !

(Zola, 1964: 1033-4)
Vizetelly’s reduction

Alas! the harvest of love was not for her!

To another were given the embraces of that husband for whose coming she herself had looked forward for so many years!

Never would she be a wife or mother; the years would come and go, and she would age in utter loneliness.
Future research: automatic (self-)censorship detection

- Automatic processing of aligned source and target texts
- Tag and parse SL and TL segments, to detect:
  - Null segment translations
  - Short TL segments (MT the SL segment, compare with TL segment, subtract common elements)
  - Sub-segment omissions (examine SL nouns and verbs, look up in thesaurus and/or ontology, look up translation results, check whether present in TL segment)
- Compile list of untranslated and mistranslated ST words/concepts
- Visualise their locations in the ST (censorship ‘hot spots’)
References


- Franzén, Nils-Olof (1958), Zola et La Joie de Vivre (Stockholm: Almqvist & Wiksell)

- Kay, Martin (1980), ‘The Proper Place of men and Machines in Language Translation’ (Palo Alto: Xerox). Available from: https://pdfs.semanticscholar.org/a9c0/4b46f2d9ff62feed73e1a0e656b6c6f28bed.pdf

- LF Aligner: http://sourceforge.net/projects/aligner/
• Massardo, Isabella, Jaap van der Meer, Maxim Khalilov (2016), TAUS Translation Technology Landscape Report https://www.taus.net/think-tank/reports

• Memsource Cloud: https://www.memsouce.com/


• TMX Editor: https://sourceforge.net/projects/tmxeditor/?source=typ_redirect