Current developments in MT

Options for the workflow in the (human) translation process

Kurt Eberle
Overview

1. Translation Architectures
2. Potential of hybrid Systems
3. Terminology preparation with rule-based MT and Corpora
4. 'Customization' in SMT extended by linguistic knowledge
5. Workflow scenarios
Translation Architectures

History

• Warren Weaver
  - 1947: letter sent to cyberneticist Norbert Wiener
    Is translation possible?
  - 1949: Memorandum Translation
    Make translation possible!
    as decoding of encoded/encrypted text, ambiguity

• Georgetown-Demonstrator IBM 1954
  - 49 simple Russian sentences
  - 'direct translation' (word-by-word, with syntactic context)
Translation Architectures

• First generation systems
  - direct translation
  - I instinctively trust the electronics boys more than I do the semanticists (Loomis)
  - 1964: ALPAC-Report

• Second generation systems
  - 1952: Bar-Hillel: Fully automatic high quality translation is not possible

Little John was looking for his toy box. Finally, he found it.
The box was in the pen. John was very happy.

- More effort, more knowledge, more linguistics!
Translation Architectures

• First generation systems
  - Direct translation
  - I instinctively trust the electronics boys more than I do the semanticists (Loomis)
  - 1964: ALPAC-Report

• Second generation systems
  - 1952: Bar-Hillel: Fully automatic high quality translation is not possible
    Little John was looking for his toy box. Finally, he found it.
    The box was in the pen. John was very happy.

Translation Architectures

- First generation systems
  - Direct translation
  - I instinctively trust the electronics boys more than I do the semanticists (Loomis)
  - 1964: ALPAC-Report

- Second generation systems
  - 1952: Bar-Hillel: Fully automatic high quality translation is not possible

  Little John was looking for his toy box. Finally, he found it.
  The box was in the pen. John was very happy.


Translation Architectures

- Second generation systems (rule-based)

- Third generation systems (statistical)
Second generation systems
(Rule-based, RBMT)

• Metal (Siemens, Langenscheidt’s T1, Lucy,..), Systran,
• Eurotra
• KANT
• Logic based Machine Translation (LMT), (McCord 1989)
  (IBM, Linguatec, Lingenio)
  – Morphological, syntactic, semantic analysis
  – Disambiguation & Transfer
  – Discourse planning, syntactic, morphologic generation
Logic based Machine Translation (LMT)

- Example: an ambiguous sentence...

Fieses Herrchen von Dackel Paul gefasst.
(Bild 15.5.11)
(nasty ..master .. of/by dachshund P. caught (arrested/bitten))
Syntactic Analysis

Transfer tree...

Restructured tree.

A nasty master taken hold by dachshund Paul.
LMT
Disambiguation by syntax

Eingabe:
:: -tsyn.
:: +all.
:: Fieses Herrchen von Dackel Paul gefasst.

Syntactic analysis no. 1. Evaluation = 3.026 ...

A nasty master taken hold by dachshund Paul.

Syntactic analysis no. 2. Evaluation = 3.1207 ...

A nasty master of dachshund Paul taken hold.
Rule-based MT (RBMT)

- Modules for analysis, transfer, generation on different levels
- costly
- many spin-offs

- Lemmatizer, PoS-Tagging,
- Syntactic and semantic text analysis
- Generation component for Q/A-systems
Translation Architectures

• Third generation systems (statistical)
  - First approaches end of 80ties, inspired by speech recognition
  - transition probability (Hidden Markov models, etc.)
  - IBM Watson Res.: Brown, Cocke, Della Pietra, Mercer, … 1989, 90, 92,…
  - ’IBM-models’
  - Uni Karlsruhe, Carnegie Mellon Alex Waibel, …
  - RWTH Aachen Hermann Ney, Franz-Josef Och, …
  - ’Feature-models’
  - Project Verbmobil
  - Google translate
Bilingual Texts → Statistical Models

... von Dackel Paul gefasst

... of (0.2), by (0.1), from (0.3), comp (0.1),...
Dachshund (0.8), idiot (0.1)
Paul (0.95),...
prepared (0.5), taken (0.3), caught (0.3), ..

Lexicon model
Alignment model, language model
(→ Source Channel Model)
Statistics-based MT (SMT)

- alignent, language models, ‘phrase tables’
- (very) cost efficient
- few(er) spin-offs
- quality?
  - depending on training sets (size + ‘balancing’)
  - combination, types, information of statistical models
  - is linguistic information helpful?

N.B.: “Every time I fire a linguist, my system’s performance improves” (Brown, IBM, TMI 1/2)
'Hybridity' becomes possible

• SMT being able to integrate linguistic information
  - from: Source-Channel-Model (Brown et al 1990)
    \[
    \hat{e}_1' = \arg\max_{e_1'} \{P(e_1') \times P(f_1'|e_1')\}
    \]
  ➢ to: Maximum-Entropy-Model (Ney/Och 2002)
    \[
    \hat{e}_1' = \arg\max_{e_1'} \{ \sum_{m=1}^{M} \lambda_m h_m(e_1', f_1') \}
    \]
    
    \[h_m \ldots \text{Linguistic Features}\]
    e.g. ... Same number of nouns, verbs
    ... agreement properties
'Hybridity' becomes possible

- SMT being able to integrate linguistic information
  - Maximum-Entropy-Model (Ney/Och 2002)
    \[
    \hat{e}_1' = \arg \max_{e_1'} \{ \sum_{m=1}^{M} \lambda_m h_m(e_1', f_1') \}
    \]
    
    \( h_m \) - Linguistic Features
    e.g. ... Same number of nouns, verbs
    ... Agreement properties
    
  ➢ More linguistics?
    - Linguistic structures instead of strings of words?
      - phrase-based SMT (Chiang 2007, ..., ISI)
      - syntax (dependency treelet)-based SMT
        (Quirk, Menezes, Cherry, 2004, ..., MSR)
Fieses Herrchen von Dackel Paul gefasst

Nasty master of dachshund Paul caught
Fieses Herrchen von Dackel Paul gefasst

Mean master of Dachshund Paul caught
Potential of hybrid Systems

- **RBMT**
  - good quality, if worked out carefully, spin-offs,
  - expensive

- **SMT**
  - Problems with 'data sparseness', 'unbalanced data'.
  - cheap

- **Combinations: 'hybrid approaches'**
  - RBMT integrating statistical methods/information from corpora
  - SMT integrating linguistic information
  - 'coarse-grained hybrid': Sequencing/parallel use of linguistic and statistical components (Eisele, Euromatrix, 2009)
Hybrid approaches
RBMT using statistical methods/information

- Learn linguistic information from corpora
  - Facts:
    Term extraction, Dictionary entries
  - Rules:
    Morphological forms
    (Inflection, Lemmata) (Eskander, Hammarström, Babych, Schütze..)
    Grammar rules (Brill, Lappin, Nivre, Bohnet,.. Malt,..)
  - Decision criteria:
    Lexical disambiguation
    (WSD) (Weaver, Yarowski, Schütze, .. SemEval)
    Weights for grammar rules
    Alternative word translations
Hybrid 'Customization'

1. SMT from TM representations instead of sent. strings (like AL)
2. using classified TMs (including linguistic classifications)

- Horizon 2020 Networking Days Luxembourg 15./16.1.14

Proposal FT2MT:
Creating SMT-Engines from TAUS-TMs
Hybrid Approaches (3)

- Coarse-grained hybrid arrangements
  (Eisele 2008, Hybrid Architectures for MT, Euromatrix)
  **Sequential and parallel use of linguistic and statistical modules**
  - Parallel use of MT systems
    Selection from results/combinations
  - Sequencing of modules (Processing Pipeline)
  Ex:
    - Standard word order for source sentence
      (‘preprocessing’)
    - SMT-System (applied to standard order)
    - ‘Reordering’ (‘postprocessing’)

23
Hybrid approaches (4)

- Human support for all types of learning
  (for RBMT, SMT data, process modules)

  ➢ 'Crowd Sourcing'
    - Web portals with user groups
    - translation suggestions etc.
    - neutral on costs for instance by exchange of information
      Ex.: TAUS – Translation memories

  ➢ Interactive translation

  ➢ Significant increase on information
Potential of hybrid approaches

- Extension and optimization of learning methods
- Designing corpora for specific purposes
- Web 2.0/3.0 – activity
  - Significant improvement of quality
- Intelligent integration into workflow: HT-MT-synchronization
  - Significant improvement on efficiency
Workflow-Szenarien

Ingredienzien

- CAT-Tool (SDL-Trados, STAR, déjà-vu, across, OmegaT,..)
- MT (Desktop/Intra/Internet-Dienst: SystranLink, LTS, Google)
- Elektronische Wörterbücher (offline/online)
- Translation memories (offline/online)
- Wort-Recherche-Tools (Daten-Harvesting)
- Erzeugung von TMs aus bilingualen Korpora
Workflow-Beispiel

- CAT-Arbeitsplatz... z.B. TRADOS

1) Text öffnen
2) Statistik TM-Anteil u. Aufwandsabschätzung
3) TM-Übersetzung
4) Zur HT/MT-Vorlage
5) automatisch übersetzen