Static and Dynamic Data in Past and Future Machine Translation

Michael Carl

CBS - CRITT
Overview

• Three origins of data-driven MT
  - concepts / representations / connectivity

• Static data-driven MT
  - example-based & statistical MT
  - representation & hybrid feature systems

• Dynamic data & MT
  - traditional translation research
  - User Activity Data (UAD) & Basic Processing Concepts (BPC)
  - Requirements for UAD query language
Conceptions of Data-driven MT

• The Translators Amanuensis (Martin Kay 1980)
   A pragmatic approach to joining man and machine

• Statistical Machine Translation (Peter F. Brown et al. 1988)
   Algorithms from the maths department

• Example-Based Machine Translation (Makato Nagao 1981)
   Mimic cognitive process of human translators
Translators Amanuensis
Martin Kay (1980)

“... an incremental approach to the problem of how machines should be used in language translation.”

“... the man and the machine are collaborating to produce not only a translation of the text but also a device whose contribution to that translation is being constantly enhanced.”

“The system will accumulate only experiences that have been agreed upon between both human and mechanical members of the team ...“
Translation Memory (TM)
Transit Editor 3.0

1. **66++** Führen Sie zur Deinstallation der TerraTVValue-Treiber das Uninstall-Utility aus. **66++**
2. **67++** Klicken Sie auf die Uninstall-Schaltfläche. **68++**
3. **69++** Nachdem alle Treiber deinstalliert wurden, erhalten Sie die Meldung, dass Uninstall erfolgreich ausgeführt wurde. **70++** Bestätigen Sie die Meldung durch Klicken auf **71++**
4. **72++** This finishes the deinstallation of the TerraTVValue drivers. **73++

**74++** Wichtige Systemeinstellungen für den Betrieb der TV-Karte **75++

**76++** Observe the following items to ensure that the TV card provides optimal performance: **77++

• **78++** The color depth for your graphics board should be set to a minimum of 16 bit or 65536 colors so that the TV image can be displayed in true-to-life color. **79++

Die Farbtiefe, mit der Ihre Graphikkarte arbeitet, können Sie über
Static & Dynamic Data in TM

• Incremental, collaborative, based on agreement

• Static data from legacy translations:
  - fuzzy match (sentence level)
  - glossaries
  - collocation tools

• Dynamic interaction during translation:
  - extend static legacy data-base
  - coarse-grained segments (sentence level)
  - coarse-grained user model

• Lacking fine-grained evaluation / exploitation of user behavior
Statistical Machine Translation
Peter F. Brown et al. (1988)

“We take the view that every sentence in one language is a possible translation of any sentence in the other language. We assign to every pair of sentences \((e, f)\) a probability \(\text{Pr}(e \mid f)\) ... the probability that a translator will produce \(e\) in the target language when presented with \(f\) in the source language.”

• Bayes' theorem provides:

\[
\text{Pr}(e \mid f) = \frac{\text{Pr}(e) \Pr(f \mid e)}{\text{Pr}(f)}
\]
Statistical Machine Translation

Peter F. Brown et al. (1993)

• Probability of source sentence $Pr(f)$ can be ignored

$$Pr(e|f) = \frac{Pr(e)Pr(f|e)}{Pr(f)}$$

• Fundamental equation in statistical Machine Translation

$$\hat{e} = \text{argmax}_e Pr(e)Pr(f|e)$$

• Toolkits available for:
  - language modelling $Pr(e)$
  - translation modelling $Pr(f|e)$
Statistical Machine Translation
Peter F. Brown et al. (1993)

“As a representation of the process by which a human being translates a passage from French to English, this equation is fanciful at best. One can hardly imagine someone rifling mentally through the list of all English passages computing the product of the a priori probability of the passage, $\Pr(e)$, and the conditional probability of the French passage given the English passage, $\Pr(f|e)$”
Example-based Machine Translation
Makoto Nagao (1981)

“Man does not translate a simple sentence by doing deep linguistic analysis, rather, [...] first, by properly decomposing an input sentence into certain fragmental phrases [...], then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence.”

- Decompose sentence into phrases
- Translate phrases into target language
- Compose phrase-translations into a sentence
Hans stellt den Klotz in der Kiste auf den Tisch.

<=>
John puts the block in the box on the table.

(Hans)_n stellt [(den Klotz)_{dp} in (der Kiste)_{dp}]_{dp} auf (den Tisch)_{dp}

<=> (John)_n puts [(the block)_{dp} in (the box)_{dp}]_{dp} on (the table)_{dp}

<=> (John)_n puts (the block)_{dp} in [(the box)_{dp} on (the table)_{dp}]_{dp}
Translation Grammar

\{n\}^1 \text{ stellen } \{dp\}^2 \text{ auf } \{dp\}^3

(art Klotz in art Kiste)_{dp}

(\{dp\}^1 \text{ in } \{dp\}^2)_n

(art Tisch)_{dp}

(art Kiste)_{dp}

(art Klotz)_{dp}

(art \{n\}^1)_{dp}

(Tisch)_n

(Kiste)_n

(Klotz)_n

(Hans)_n

<=> \{n\}^1 \text{ put } \{dp\}^2 \text{ on } \{dp\}^3

<=> (the block in the box)_{dp}

<=> (\{dp\}^1 \text{ in } \{dp\}^2)_n

<=> (the table)_{dp}

<=> (the box)_{dp}

<=> (the block)_{dp}

<=> (the \{n\}^1)_{dp}

<=> (table)_n

<=> (box)_n

<=> (block)_n

<=> (John)_n
Data-Oriented Translation

just fell  \(<-->\) vient de tomber

Finite verbs „fell“ and „tomber“ are not translational equivalents
Relaxing Constraints in LFG-DOT

- Relax TENSE and FIN features
- `<FALL, TOMBER>` can be linked
Complexity of Connectivity

- Combining recursive structures
  - exponential
- Linking feature sub-systems
  - exponential
- Disambiguating
  - readings & meanings
  - segmentation

- How to choose appropriate prolongation of structures?
  - Intuitive modelling of feature constraints:
    rule-based constraint-formalisms no resort
Statistical Machine Translation
Hermann Ney (2005)

Statistical Machine Translation investigates:

„the more or less purely algorithmic concepts of how we model the dependencies of the data.“

- Select appropriate features
- Train functions on a learning corpus
- Apply functions to search best translation
Hybrid Machine Translation

- Generalization of Noisy Channel Model allows combination of different, heterogeneous sub-systems $h$:

$$\hat{e} = \arg\max \sum_{i=1}^{M} w_i h_i()$$

- $h_i$ Feature function
- $w_i$ Weight of feature function

- Automatic Evaluation Scores
  - BLEU, NIST, etc.
Translation Hypotheses AND/OR Graph for:

**Hans kommt nicht**

\{lu=Hans,c=noun, wnr=1\}
  @ \{c=noun}@\{lu=hans,c=NP0\}..

,\{lu=nicht,c=adv,wnr=3\}
  @ \{c=verb}@\{lu=do,c=VDZ\},\{lu=not,c=XX0\}.
  ; \{c=adv}@\{lu=not,c=XX0\}..

,\{lu=kommen,c=verb,wnr=2\}
  @ \{c=verb}@\{lu=come,c=VVB\}.
  ; \{c=verb}@\{lu=come,c=VVB\},\{lu=along,c=AVP\}.
  ; \{c=verb}@\{lu=come,c=VVB\},\{lu=off,c=AVP\}.
  ; \{c=verb}@\{lu=come,c=VVB\},\{lu=up,c=AVP\}..
Scoring n-best Translations

- Traverse AND/OR graph to score n-best Translations
- Breadth first search (Beam-search algorithm)
- Feature Function:
  - Lemma Language Model (3-gram, 4-gram)
  - Tag Language Model (5-gram to 7-gram)
  - Lemma/tag co-occurrence model
- Combination of feature functions Log-linear
Output

<table>
<thead>
<tr>
<th>lemma,</th>
<th>tag,</th>
<th>#dico,</th>
<th>expander rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>AT0</td>
<td>146471</td>
<td></td>
</tr>
<tr>
<td>company</td>
<td>NN1</td>
<td>268244</td>
<td></td>
</tr>
<tr>
<td>is</td>
<td>VBD</td>
<td>604071</td>
<td>PermFinVerb_hs</td>
</tr>
<tr>
<td>buy</td>
<td>VVN</td>
<td>307263</td>
<td>PermFinVerb_hs</td>
</tr>
<tr>
<td>by</td>
<td>PRP</td>
<td>587268</td>
<td>PermFinVerb_hs</td>
</tr>
<tr>
<td>hans</td>
<td>NP0</td>
<td>265524</td>
<td>PermFinVerb_hs</td>
</tr>
<tr>
<td>.</td>
<td>PUN</td>
<td>367491</td>
<td></td>
</tr>
</tbody>
</table>

<s id=3-0 lp="-9.227912">
the   AT0  146471
company NN1 268244
is     VBD  604071    PermFinVerb_hs
buy    VVN  307263    PermFinVerb_hs
by     PRP  587268    PermFinVerb_hs
hans   NP0  265524    PermFinVerb_hs
.      PUN  367491    
</s>
Dependency Treelet Translation
Quirk & Menezes (2006)

• Resources:
  - (shallow) source-language dependency parser
  - target language word segmentation
  - unsupervised word alignment

• Learn treelet translations
  - arbitrary connected subgraph of aligned dependency trees

• Project source tree onto the target sentences
  - extension of tree-to-string translation

• Train statistical models on aligned dependency tree corpus
Hybrid Feature Integration

• Decoding depends on
  - $S$: source dependency tree
  - $T$: target dependency tree
  - $A$: word alignment between the source and target trees
  - $I$: set of treelet partitioning $S$ and $T$ into treelets

• Find translation which maximises:

$$SCORE(A,T,A,I) = \sum_{f \in F} \log f(S,T,A,I)$$
Static Data-driven MT

- Use corpora and examples to train:
  - decomposition operations
  - translation relations
  - composition operations

- Combine feature functions to integrate heterogeneous sub-systems

- No user modelling
- No collaboration between user & MT system
- No targeted translation
- No high quality translations
Dynamic Data and MT

- Martin Kay (1980): “... man and the machine are collaborating to produce [...] a translation ...”
- Makoto Nagao (1981): “Man does not translate [...] by doing deep linguistic analysis ...”

But: how does Man translate?

- Traditional empirical translation research techniques
- TRANSLOG: recording keystrokes
- User-Activity Data:
  - recording eye-movement and keystroke behavior
- Uncover Basic Processing Concepts (BPC)
  - building blocks of mental representation
Think Aloud Protocol (TAP)
Research into Translation Processes

• View translation as a decision making process:
  – establish complex inventory (Lörscher, Krings)
    • strategies performed by translators
    • meaning operations

• Processing is disturbed:
  – delay of translation by 25%
  – degenerative effect on segmentation and translation rhythm
TRANSLOG
Recording Keystrokes in Time

• Temporal patterns reflect cognitive rhythm
• Different in monolingual text production & text translation:
  - Hierarchical structure of pauses between segments
  - Translation rhythm does not reflect linguistic structure

• Peculiarities of translation production:
  - Translators do not think about sentence/paragraph planning
  - Fluent translation is disturbed by local problems
    • Unpredictable structure, semantic problems
User Activity Data (UAD)

Eye-movement & Keystroke activities

• Eye movement depends on:
  - length/ambiguity of words
  - probability of occurrence
  - familiarity with specific words and concepts

• Multiple fixations within a word and/or returning refixation(s) indicate:
  - failure of successful meaning construction
  - failure of mapping meaning into target language

• Regressive saccades to reinspect failed meaning construction
Spielberg shows Beijing red card over Darfur

In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China’s backing for Sudan’s policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government. China, which has extensive investments in the Sudanese oil industry, maintains close links with the Government, which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities, Spielberg maintains that the international community, and particularly China, should do more to end the suffering.

Spielberg giver Beijing det røde kort for Darfur
I en der helt sikkert ryster den kinesiske regering, trak Steven Spielberg sig fra de olympiske lege i Beijing for at protestere imod Kinas støtte af den sudanske...i Darfur. Hans tilbagetrækning kommer i ....kampene i Darfu.
Spielberg shows Beijing red card over Darfur

In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China’s backing for Sudan’s policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government. China, which has extensive investments in the Sudanese oil industry, maintains close links with the Government, which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities, Spielberg maintains that the international community, and particularly China, should do more to end the suffering.

Spielberg giver Beijing det røde kort for Darfur

I en der helt sikkert rystet den kinesiske regering, trak Steven Spielberg sig fra de olympiske lege i Beijing for at protestere imod Kinas støtte af den sudanske...i Darfur. Hans tilbage trækning kommer i ....kampene i Darfur. Kina som har en del investeringer i den sudaneske olieindustri, har tætte bånd med regeringen, heunder en minister der er blevet anklaget for ...af den Internationale domstol i Haag. |
UAD and Basic Processing Concepts

• Basic Processing Concepts (BPC):
  - link functional features of action and sensory input
  - building blocks of mental representation

• Infer BPC from User-Activity Data (UAD):
  - sensory input: eye-movements
    • reading and construction of source text meaning
  - actions: keyboard activity
    • discharge of information stored in working memory

• BPC provide detailed picture of processing for:
  - constructing meaning during reading
  - mapping/modification of target representation
BPCs for Postediting

• Detect from eye-movements & background knowledge whether translation is:
  - wrong, awkward, confusing,
    conform to cooperate or personal style

• Detect from keyboard activities:
  - Linguistic operations:
    change of POS, adjust agreement, insert/delete words, ...

• Infer aims of modification:
  - increase fluency or coherence, remove ambiguities, add information, reduce complexity, change focus, clarify relation, ...
Uncover BPC in UAD

- Develop query language to detect dependencies between:
  - eye-movement (construction of meaning)
  - keyboard activities (discharge/arrangement of information)
  - properties of source text/translation

- Elaborate 'clean' manually-corrected reference data:
  - re-adjust gaze-to-word mapping
  - assign linguistic information

- GWM-remapper:
  - visualise activity patterns
    - keyboard, samples, fixations, mappings
  - correct fixations & mapping data
  - store corrected data
Spielberg shows Beijing red card over Darfur

In a gesture sure to rattle the Chinese Government Steven Spielberg pulled out of the Beijing Olympics to protest against China’s backing for Sudan’s policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China which has sought to halt the negative fallout from having close ties to the Sudanese government. China which has extensive investments in the Sudanese oil industry maintains close links with the Government which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities Spielberg maintains that the international community and particularly China should do more to end the suffering.
Spielberg shows Beijing red card over Darfur

In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China's backing for Sudan's policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government. China, which has extensive investments in the Sudanese oil industry, maintains close links with the Government, which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities, Spielberg maintains that the international community and particularly China should do more to end the suffering.
Spielberg shows Beijing red card over Darfur

In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China's backing for Sudan's policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China which has sought to halt the negative fallout from having close ties to the Sudanese government China which has extensive investments in the Sudanese oil industry maintains close links with the Government which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities Spielberg maintains that the international community and particularly China should do more to end the suffering.
Spielberg shows Beijing red card over Darfur

In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China's backing for Sudan's policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China which has sought to halt the negative fallout from having close ties to the Sudanese government. China, which has extensive investments in the Sudanese oil industry, maintains close links with the Government which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities, Spielberg maintains that the international community and particularly China should do more to end the suffering.
Spielberg shows Beijing red card over Darfur

In a gesture sure to rattle the Chinese Government, Steven Spielberg pulled out of the Beijing Olympics to protest against China’s backing for Sudan’s policy in Darfur. His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government. China, which has extensive investments in the Sudanese oil industry, maintains close links with the Government, which includes one minister charged with crimes against humanity by the International Criminal Court in The Hague. Although emphasizing that Khartoum bears the bulk of the responsibility for these ongoing atrocities, Spielberg maintains that the international community and particularly China should do more to end the suffering.
Conclusion

• To date, data-driven MT is:
  – hybrid, static

• New research method for studying dynamic human activities during reading and post-editing:
  – uncover patterns of UAD (eye-movement, keystroke)
  – detect dependencies in UAD and properties of text
  – determine Basic Processing Concepts (BPC)
  – express BPC in terms of features

=> fine-grained model of posteditor/user

• Ultimately: feed-back BPC into MT