A Data-Driven Approach to Deep Machine Translation

Michael Jellinghaus
Saarland University
micha@coli.uni-sb.de
Overview

- Motivation
  - Characterisation of statistical and transfer-based MT
  - Hybrid system idea
- Automatic acquisition of transfer rules
  - Workflow
  - Example
  - Some details
- Evaluation
- Outlook
Statistical Machine Translation

- Quick to develop
  - Translation model learned from parallel corpora
  - Target language model learned from monolingual corpora

- High coverage
  - Covers all technical terms etc. if seen in training data
    - e.g. *Steueroase / paradis fiscal ➔ tax haven*
  - Robust: always delivers some output

but...
Problems with syntactically or semantically more complex input
(examples from Google Translate):

Der von Browne gejagte Hund bellte.
(R: The dog chased by Browne barked.)

- The Hunted Browne dog barked. (March 2008)
- The Browne gejagte dog barked. (May 2008)

Der von der Katze gejagte Hund bellte.
(R: The dog chased by the cat barked.)

- The cat Hunted by the dog barked. (March 2008)
- The cat gejagte the dog barked. (May 2008)
Problems with syntactically or semantically more complex input
(examples from Google Translate):

Abrams versprach Browne zu bellen.
(R: Abrams promised Browne to bark.)
→ Abrams Browne promised to bark. (March 2008)
→ Abrams promised Browne to bark. (May 2008)

Michael versprach Georg zu bellen.
(R: Michael promised Georg to bark.)
→ George Michael promised to bark. (May 2008)
Deep Transfer-Based Machine Translation

- LOGON project

HPSG/LFG analysis

transfer at semantic level

HPSG generation

source text

target text

2008-05-16
Michael Jellinghaus
Deep Transfer-Based Machine Translation

Minimal Recursion Semantics example

Der Hund jagt die Katze. (The dog chases the cat.)

[ LTOP: h1
  INDEX: e2 [ e MOOD: INDICATIVE TENSE: PRESENT ]
  RELS: <
    [ "_def_q_rel"
      LBL: h3
      ARG0: x5 [ x PERS: 3 NUM: SG ]
      RSTR: h4
      BODY: h6 ]
    [ "_hund_n_rel"
      LBL: h7
      ARG0: x5 ]
    [ "_def_q_rel"
      LBL: h10
      ARG0: x9
      RSTR: h11
      BODY: h12 ]
    [ "_jagen_v_rel"
      LBL: h8
      ARG0: e2
      ARG1: x5
      ARG2: x9 [ x PERS: 3 NUM: SG ] ]
    [ "_katze_n_rel"
      LBL: h13
      ARG0: x9 ]
    [ prop-or-ques_m_rel
      LBL: h1
      ARG0: e2
      MARG: h14
      TPC: x5 ] >

HCONS: < h14 qeq h8 h4 qeq h7 h11 qeq h13 > ]
Deep Transfer-Based Machine Translation

- **Advantages**
  - Preserves meaning
  - Grammatical output

- **Disadvantages**
  - High development cost due to manual rule production
  - Weak on idiomaticity, e.g. *paradis fiscal* → *fiscal paradise*
  - Low coverage, e.g. *Steueroase* probably not in lexicon
### Complementary Approaches to MT

- **Idea:** Combine advantages by learning transfer rules from parallel corpora

<table>
<thead>
<tr>
<th>Feature</th>
<th>SMT</th>
<th>DTBMT</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>development speed</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>grammaticality</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>lexical semantics</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>structural semantics</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>coverage</td>
<td>+</td>
<td>-</td>
<td>-(?)</td>
</tr>
</tbody>
</table>
Overview

- **Motivation**
  - Project context
  - Characterisation of statistical and transfer-based MT
  - Hybrid system idea

- **Automatic acquisition of transfer rules**
  - Workflow
  - Example
  - Some details

- **Evaluation**

- **Outlook**
Transfer Rule Acquisition Workflow

- sentence-aligned parallel corpus

- parse
  - parallel MRS corpus

- align etc.
  - word alignment information etc.

- semantic alignment & rule extraction

- set of transfer rules
Minimal Recursion Semantics example

*Der Hund jagt die Katze. (The dog chases the cat.)*

```
[ LTOP: h1
  INDEX: e2 [ e MOOD: INDICATIVE TENSE: PRESENT ]
  RELS: <
    [ "_def_q_rel"
      LBL: h3
      ARG0: x5 [ x PERS: 3 NUM: SG ]
      RSTR: h4
      BODY: h6 ]
    [ "_jagen_v_rel"
      LBL: h8
      ARG0: e2
      ARG1: x5
      ARG2: x9 [ x PERS: 3 NUM: SG ] ]
    [ "_hund_n_rel"
      LBL: h7
      ARG0: x5 ]
    [ "_katze_n_rel"
      LBL: h10
      ARG0: x9 ]
  ]
[ prop-or-ques_m_rel
  LBL: h1
  ARG0: e2
  MARG: h14
  TPC: x5 ]>
HCONS: < h14 qeq h8 h4 qeq h7 h11 qeq h13 > ]
```
Transfer Rule Acquisition Example

prop-or-ques_m_rel

"_jagen_v_rel"

"_def_q_rel"

"_hund_n_rel"

X

prop-or-ques_m_rel

"_chase_v_1_rel"

"_def_q_rel"

"_katze_n_rel"

X

"_dog_n_1_rel"

"_cat_n_1_rel"

X

dog_rule_0 := monotonic_omtr &
[ INPUT [ RELS < [ PRED "_dog_n_1_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ] > ] ],
OUTPUT [ RELS < [ PRED "_hund_n_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ] > ] ].
Transfer Rule Acquisition Example

```
prop-or-ques_m_rel
```

```
"_jagen_v_rel"
```

```
"_def_q_rel"
```

```
"_hund_n_rel"
```

```
"_katze_n_rel"
```

```
prop-or-ques_m_rel
```

```
"_chase_v_1_rel"
```

```
"_the_q_rel"
```

```
"_def_q_rel"
```

```
"_dog_n_1_rel"
```

```
"_cat_n_1_rel"
```

```
the_rule_0 := monotonic_omtr &
```

```
[ INPUT [ RELS < [ PRED _the_q_rel, LBL #1, RSTR #2, ARG0 #3 & [PERS #4, NUM #5], BODY #6 ] > ],
```

```
OUTPUT [ RELS < [ PRED "_def_q_rel", LBL #1, RSTR #2, ARG0 #3 & [PERS #4, NUM #5], BODY #6 ] > ] ].
```

Transfer Rule Acquisition Example

```
prop-or-ques_m_rel
   "_jagen_v_rel"
   "_def_q_rel"
   "_hund_n_rel"

prop-or-ques_m_rel
   "_chase_v_1_rel"
   "_def_q_rel"
   "_dog_n_1_rel"

"_katze_n_rel"

"_cat_n_1_rel"

the_rule_1 := monotonic_omtr &
[ INPUT [ RELS < [ PRED "_dog_n_1_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ],
  [ PRED _the_q_rel, LBL #5, RSTR #6, ARG0 #2, BODY #7 ] >,
  HCONS < [qeq & HARG #6, LARG #1] > ],
 OUTPUT [ RELS < [ PRED "_def_q_rel", LBL #5, RSTR #6, ARG0 #2 & [PERS #3, NUM #4],
    BODY #7 ],
  [ PRED "_hund_n_rel", LBL #1, ARG0 #2 ] >,
  HCONS < [qeq & HARG #6, LARG #1] > ] ].
```
Transfer Rule Acquisition Example

\[
\text{cat\_rule\_0 := monotonic\_omtr} \& \left[\text{INPUT [ RELS < [ PRED "\_cat\_n\_1\_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ] > ]}, \right. \\
\left. \text{OUTPUT [ RELS < [ PRED "\_katze\_n\_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ] > ]}\right].
\]
chase_rule_0 := monotonic.omtr &
[ INPUT [ RELS < [ PRED "_chase_v_1_rel", LBL #1, ARG0 #2 & [MOOD #3, TENSE #4],
ARG2 #5, ARG1 #6 ] > ],
OUTPUT [ RELS < [ PRED "_jagen_v_rel", LBL #1, ARG0 #2 & [TENSE #4, MOOD #3],
ARG2 #5, ARG1 #6 ] > ] ].

Transfer Rule Acquisition Example
chase_rule_1 := monotonic_omtr &
[ INPUT [ RELS < [ PRED "_dog_n_1_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ],
  [ PRED _the_q_rel, LBL #5, RSTR #6, ARG0 #2, BODY #7 ],
  [ PRED ",_chase_v_1_rel", LBL #8, ARG0 #9 & [MOOD #10, TENSE #11], ARG2 #8, ARG1 #2 ] >,
  HCONS < [qeq & HARG #6, LARG #1] > ],
OUTPUT [ RELS < [ PRED "_jagen_v_rel", LBL #8, ARG0 #9 & [TENSE #11, MOOD #10],
  ARG2 #8, ARG1 #2 & [PERS #3, NUM #4] ],
  [ PRED "_def_q_rel", LBL #5, RSTR #6, ARG0 #2, BODY #7 ],
  [ PRED ",_hund_n_rel", LBL #1, ARG0 #2 ] >,
  HCONS < [qeq & HARG #6, LARG #1] > ] ].
prop-or-ques_m_rel

"jagen_v_rel"

"_def_q_rel"

"_hund_n_rel"

"_katze_n_rel"

prop-or-ques_m_rel

"chase_v_1_rel"

_the_q_rel

"_dog_n_1_rel"

"_cat_n_1_rel"

chase_rule_2 := monotonic.omtr &

[ INPUT [ RELS < [ PRED "_the_q_rel", LBL #1, RSTR #2, ARG0 #3 & [PERS #4, NUM #5], BODY #6 ],
    [ PRED "_chase_v_1_rel", LBL #7, ARG0 #8 & [MOOD #9, TENSE #10], ARG2 #3, ARG1 #11 ],
    [ PRED "_cat_n_1_rel", LBL #12, ARG0 #3 ] >,
    HCONS < [qeq & HARG #2, LARG #12] > ],
OUTPUT [ RELS < [ PRED "_jagen_v_rel", LBL #7, ARG0 #8 & [TENSE #10, MOOD #9],
    ARG2 #3 & [PERS #4, NUM #5], ARG1 #11 ],
    [ PRED "_def_q_rel", LBL #1, RSTR #2, ARG0 #3, BODY #6 ],
    [ PRED "_katze_n_rel", LBL #12, ARG0 #3 ] >,
    HCONS < [qeq & HARG #2, LARG #12] > ]].
Kinds of Transfer Rules

- **Simple “lexical” rules**
- **Rules with multiple EPs on input and/or output side**
  - Multi-word expressions / compounds
  - Phrasal translations
    - e.g., *the book I like most* vs. *my favourite book*
  - EPs together with one or more of their argument “subtrees”, e.g.,
    - *the cat eats ... ➞ die Katze frisst ... (not isst)*
    - ... *sitzt auf der Bank ➞ ... sits on the bench (not bank)*
- But neither complete sentences nor less interesting collocations such as verb-adjective combinations etc.
Transfer Rule Acquisition Pipeline

- Preprocessing
  - Tokenization
  - Part-of-speech tagging
  - Named entity recognition
- Parsing
- Treebanking
  - parse selection (done manually in experiments)
    - Example for ambiguity: Das Kind jagt die Katze
- Semantic alignment and rule extraction
  - Algorithm is language-independent
- Construction of transfer rule set
Construction of Transfer Rule Set

- **Quality control**
  - Learned rules are rejected unless complete alignment achieved

- **Internal order of rule set:**
  - Sort rules by number of input EPs ("specific rules first" strategy for increased idiomaticity)
  - Then sort by rules' extraction frequency (in order to eliminate noise)

- **Examples of noise:**
  - *Wer... → what group...* (loose translation)
  - *Das Kind jagt die Katze* (ambiguity)
  - Other errors at the various levels (parsing, alignment, ...)
Overview

Motivation
- Project context
- Characterisation of statistical and transfer-based MT
- Hybrid system idea

Automatic acquisition of transfer rules
- Workflow
- Example
- Some details

Evaluation

Outlook
Evaluation

- Closed evaluation on MRS Test Suite (107 sentences)
  - All sentences contributing to rule set were translated correctly (plus additional results due to ambiguity or syntactic variation from the generator)
- Evaluation on unseen data (but lexical items and constructions had been seen; 79 sentences)
  - As above, except for 2 sentences that were translated incorrectly (could be tracked to treebanking error)
- Evaluation on CLEF corpus (QA data; about 1600 sentences)
  - No clean results yet :(
Overview

Motivation
- Project context
- Characterisation of statistical and transfer-based MT
- Hybrid system idea

Automatic acquisition of transfer rules
- Workflow
- Example
- Some details

Evaluation

Outlook
Outlook

- Evaluation on larger corpora and other languages
  - Quantitative
    - Coverage, BLEU score etc.
  - Qualitative
    - Which phenomena are difficult/easy?
    - Compare with SMT (“division of labour”)
- Automatic parse selection
  - Goal: eliminate manual treebanking step
- Build hybrid systems
  - Back-off to SMT when out of coverage
  - Provide high-confidence phrase pairs to SMT phrase table
Outlook

- Rule set experiments
  - Maximum size?
  - What are interesting collocations?
  - Generalise rules
    - HPSG types
    - Semantic classes (information from ontologies)
  - Stochastification(?)

- Learn more rules:
  - Extract at least phrase translation rules if sentences cannot be aligned completely
  - Acquire rules from dictionaries etc.

- Use in application-based evaluation
One Last Example from Google Translate ...

Input:
Danke, dass ihr meinem Talk so aufmerksam gefolgt seid ohne einzuschlafen
Zuletzt nehme ich noch gerne eure Fragen und Anmerkungen entgegen

Output:
Thank you, that you talk to my attention are followed without einzuschlafen
Recently, I still like your questions and comments contrary to