Highlighting Matched and Mismatched Segments in Translation Memory Output through Sub-Tree Alignment

Ventsislav Zhechev
Outline

- Translation Memory Backend
- Sub-Tree Alignment
- Translation-Alignment Algorithm
- Evaluating the usefulness of Statistical Machine Translation
- Future Work
- Conclusion
Translation Memory Backend

- A PostgreSQL database containing the plain TM data
- Can perform fuzzy matching based on a fast character-based Levenstein-distance search
- The Levenstein-based distance of the fuzzy match is normalised by the number of characters in the shorter sentence
- Integrate with a proper TM in the future
Sub-Tree Alignment

- Main use: for generating training resources for Syntax-Based Machine Translation
  - i.e. Parallel Treebanks
I do not think it is necessary for classic cars to be part of the directive.
I am not looking for such rigidly high recycling quotas when it comes to special-purpose vehicles either.
I want special-purpose vehicles such as ambulances to have high recovery quotas.
This is my main concern in this matter.

Ich halte es nicht für notwendig, daß Oldtimer Bestandteil dieser Richtlinie sind.
Auch bei Sonderfahrzeugen strebe ich nicht so unbedingt hohe Recyclingquoten an.
Ich habe den Wunsch, daß Sonderfahrzeuge wie Krankenwagen hohe Rettungsquoten haben.
Das ist meine Hauptsorge in diesem Bereich.
This is my main concern in this matter.

Das ist meine Hauptsorge in diesem Bereich.
This is my main concern in this matter.
This is my main concern in this matter.
The sub-tree aligner operates on parsed data

For many languages no parsers are available

- Retraining existing parsers for new languages may require significant resources

- The string-to-string aligner operates on plain sentences
Alignment Algorithm

Bilingual Alignment

- Align the SL fuzzy match to its TL translation from the TM
- The sub-tree aligner operates on plain unparsed data
- The probabilistic bilingual dictionary it uses may be generated using an off-the-shelf word-alignment tool (eg. GIZA++)
Alignment Algorithm

Bilingual Alignment

M

T

X1 sender
X2 ’s
X3 email
X4 address
X5 .
X1 adresse
X2 électronique
X3 de
X4 l’
X5 expéditeur
X6 du
X7 message
X8 .
Alignment Algorithm
Bilingual Alignment

X1: sender
X2: 's
X3: email
X4: address
X5: .
X1: adresse électronique
X2: de
X3: l’expéditeur
X4: du message
X5
X6
X7
X8
Alignment Algorithm

Bilingual Alignment
Alignment Algorithm

Monolingual Alignment

- Align the SL TM fuzzy match to the input sentence

- Namely, the plain input sentence to the structure derived for the SL TM fuzzy match during the bilingual alignment

- Use a dummy probabilistic dictionary, where each SL word available in the TM is aligned to itself with probability 1.
Alignment Algorithm
Monolingual Alignment
Alignment Algorithm
Monolingual Alignment
Alignment Algorithm
Monolingual Alignment
Alignment Algorithm

Matching

* The structure of the SL TM sentence is used as a pivot to align the structures of the input sentence and the TL TM sentence.
Alignment Algorithm

Matching
Alignment Algorithm

Matching

X1 sender
X2 email
X3 address
X4 X18 X6 du
X5 expéditeur
X6 message
de l'
X8 .
exécuter
Alignment Algorithm

Matching

X1 sender
X2 email
X3 address

X1 adresse
X2 électronique
X3 de l’expéditeur
Alignment Algorithm
Matching
Alignment Algorithm

Matching
SMT Backend

- Use standard Moses for phrase-based SMT

- Two modes of operation:
  - *comb* translate the mismatched parts of the input individually using the SMT backend
  - *xml* mark-up the matched parts of the input with their translations and translate the marked-up input as a whole
Reordering

- Use the parallel treebank to reorder the SL side of the TM to conform to the TL word order
- The SMT backend is then retrained to generate a ‘reordered’ model
- Both the regular and ‘reordered’ models are used during translation
Evaluation Data

- Symantec EN–FR training data
  - 108,953 segment pairs
  - 13.2 EN average length
    15.0 FR average length
  - 41,379 EN unique tokens
    49,971 FR unique tokens

- Symantec EN–FR test data
  - 4,977 segment pairs
  - 9.2 EN average length
    10.9 FR average length
Evaluation Data

- Large number of XML tags
  - 2,049 EN unique tags
  - 2,653 FR unique tags

- Many ‘special’ strings
  - File paths
  - URLs
  - e-mail addresses
  - RTF formatting
  - XML tags with translatable parameters

- Meta-tag handling tool
- Specialised tokenizer
Evaluation Results

![Graphs showing BLEU, METEOR, TER, and Inverse F-Score results across different FMS range/segments for different methods (tm, comb, xml, direct).]
Evaluation Results
Evaluation Results

Inverse F-Score

FMS Range/Segments

0…50/1963 50…60/779 60…70/621 70…80/537 80…90/537 90…100/375 100/165

0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8

tm
comb
xml
direct
Evaluation Results

![Graph showing the relationship between FMS Range/Segments and Length.](image)

- **FMS Range/Segments**: 0…50/1963, 50…60/779, 60…70/621, 70…80/537, 80…90/537, 90…100/375, 100/165
- **Length**: The length values range from 2 to 18, decreasing at higher FMS ranges and segments, with a dip around 70…80/537, followed by an increase around 90…100/375 and a significant decrease towards 100/165.
Future Work

- Develop a prototype implementation of the presented work
- Integrate this framework with a proper TM system
- Perform a user study to evaluate the effect of this framework on post-editing speed
- Further develop the meta-tag handling tool
  - possibly integrating it with the alignment and SMT backends
- Improve the reordering accuracy
- Run experiments where the SMT backend has been trained on additional data, besides the TM
Thank you!

http://VentsislavZhechev.eu

contact@VentsislavZhechev.eu