Predicting MT Fluency from IE Precision and Recall

Tony Hartley, Brighton, UK
Martin Rajman, EPFL, CH
ISLE Keywords

• Quality of output text as a whole
  – Fluency (itself as a predictor of fidelity)
  – Utility of output for IE
• Automation
• F-score, Precision, Recall,
Data Sources

• Fluency
  – DARPA 94 scores for French => English
  – 100 source texts * (1 HT + 5 MT)

• IE
  – Output from SRI Highlight engine
    – http://www-cgi.cam.sri.com/highlight/
Data Sampling

• Select 3 ‘systems’
  – Human expert -- mean fluency: 0.851934
  – S1 -- highest mean fluency: 0.508431
  – S2 -- lowest mean fluency: 0.124882

• Cover ‘best to worse’ for each
  – Aim -- select 20 texts
  – Actual -- 17 texts
IE Task

Highlight

Intelligent Search and Extraction of Information

Choose options below and press analyse for a live demo.

What would you like to search for?
- losses by gas or oil companies
- new projects, ventures, or bids concerning gas or oil companies
- relationships between individuals and companies
- actions by individuals, objects, companies

Which text/texts would you like to search?
- Your own text file. Please enter a URL
  http://
For its part, a few weeks ago, Boeing announced 226 orders for the year, but that 214 aircraft would be retired.

Nevertheless, deliveries have maintained a satisfactory pace for the European builder. They should rise to 138 in 1993, compared to 157 in 1992, which was a very good year.

Nevertheless, 1993 sales should be higher than in 1992 (7.6 billion dollars), taking into account the delivery of large transport long range A 340s.

As the chief cause for celebration, 1993 will have been marked by the launching of the small transport Airbus A 319 at the Salon du Bourget.

The A 319 was put into service in March 1993 with Lufthansa and Air France, and the first A 330 was delivered to its launching company, Air Inter, in December.

<table>
<thead>
<tr>
<th>entity</th>
<th>relation</th>
<th>entity</th>
<th>link to text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airbus Industrie</td>
<td>Recorded</td>
<td>59 Cancelled Orders</td>
<td>source</td>
</tr>
<tr>
<td>Airbus Industrie</td>
<td>issue</td>
<td>its profit and loss picture</td>
<td>source</td>
</tr>
<tr>
<td>No official reason</td>
<td>be - given</td>
<td>-</td>
<td>source</td>
</tr>
</tbody>
</table>
IE Scoring

- Manual, but using simple, objective rules
- ‘Gold Standard’ given by expert translation
  - number of cells filled
- ‘Strict’ match = identity
- ‘Relaxed’ match = near miss
  - quarter :: third school quarter
  - hoped :: wished
  - Airbus Industrie :: Airbus industry
Precision and Recall

• Precision $P$
  – (correct cells / actual cells)

• Recall $R$
  – (actual cells / reference cells)

• F-score
  – $2 \times P \times R / (P + R)$
Correlations (not): F-score

S1 (Fstrict:Fluency)

S2 (Fstrict:Fluency)

S1 (Frel:Fluency)

S2 (Frel:Fluency)
Correlations (not): Precision

S1 (Pscript:Fluency)

S2 (Pscript:Fluency)

S1 (Pre:Fluency)

S2 (Pre:Fluency)
Correlations (not): Recall

S1 (Rstrict:Fluency)

S2 (Rstrict:Fluency)

S1 (Rrel:Fluency)

S2 (Rrel:Fluency)
Conclusions

- A significant correlation is observed for S2 (for precision and recall)
- The same is not true for S1!
- For S1, the correlation is higher for the F score than for P and R scores
Observations on translation data

• MT yields structure preservation by default
• Does ‘free-ness’ of expert translation tarnish Gold Standard?
  – Finite verb in ST => non-finite verb in TT (IE under-retrieves)
  – Nominalisation in ST => finite verb in TT (IE over-retrieves)
Observations on IE data

• Poor performance of *Highlight*
  – particularly on expert translations?
• Long distances between Entities and Relator
  – post-modification of NP
  – parenthetical material
Next Steps

• Correlate with Adequacy rather than Fluency?
• Use whole data set (S2 bad choice?)
• Focus on extracted doubles and triples rather than single cells?
• Use another IE engine?
• Use ST to establish Gold Standard?