From Statistical Term Extraction to Hybrid Machine Translation

Petra Wolf, Ulrike Bernardi
Lucy Software and Services, Munich

Christian Federmann, Sabine Hunsicker
DFKI, Saarbrücken
Contents

- Motivation

- Intelligent Terminology Extraction for Hybrid MT

- Evaluation

- Conclusion
Motivation

- SMT: superior selection of target entities
- Phrase table access at RBMT transfer: strong & weak
- Deeper intertwined hybrid extension

\[ \Rightarrow \text{LiSTE}X: \]
Hybrid Transfer by
Linguistically augmented Statistical Terminology EXtraction
Intelligent Terminology Extraction

Parallel Documents

Ling. Processing:
Tokenization
POS Tagging
Lemmatization

Creation of RBMT Trees

Statistical Word Alignment

Initial Term Pairs
**Initial Term Pairs - Examples**

Abfallverbrennungsanlage; NST; waste incineration plant; NST;3; Die Technologie zur Rauchgasreinigung in Abfallverbrennungsanlagen existiert bereits.; The technology for cleaning up fumes from waste incineration plants already exists.; NST AST; NST NST

Beitrittsverhandlung; NST; accession negotiation; NST;208; Malta tritt in die Beitrittsverhandlungen ein.; Malta is entering the accession negotiations.; NST NST NST
Intelligent Terminology Extraction

Parallel Documents

Ling. Processing:
- Tokenization
- POS Tagging
- Lemmatization

Creation of RBMT Trees

Statistical Word Alignment

Initial Term Pairs

Ling. Filtering & Transformation:
- Bidirectional Reconciliation
- Collocation Check
- Ling. Driven Data Splitting
Collocation Check

Reduction of false entries:

- Extract frequencies (for single and multiwords)
- Reject single terms with lower/equal frequency than corresponding multiwords

Tour de France  →  Tour de France  -  Tour de France
              →  France  -  France
Ling. Driven Data Splitting

Higher precision by data splitting in case of:

- Lemmatizer does not find correct lemma and just guesses it
- Special characters, e.g. punctuation or integers
- Category changes

Quality Splitting:
1. Single words on source as well as on target side
2. Single words as source, multiword expressions as targets
3. Multiword expressions as source, single words as targets
4. Multiword expressions on source as well as on target side

- POS distinction for nouns, verbs, adjectives, adverbs
Intelligent Terminology Extraction

Parallel Documents

Ling. Processing:
- Tokenization
- POS Tagging
- Lemmatization

Creation of RBMT Trees

Statistical Word Alignment

Initial Term Pairs

Ling. Filtering & Transformation:
- Bidirectional Reconciliation
- Collocation Check
- Ling. Driven Data Splitting

Final Term Pairs

Ling. Selection and Feature Value Augmentation
Ling. Selection & Feature Value Augmentation

- **Quality Precision & Selection:**
  - Resolve lemmatizer and derivational errors
  - Rectify wrong or missing endings
  - Delete senseless terms, such as *of of*
  - Delete wrongly categorized entries
  - Evaluate nouns in plural form
  - Delete wrong translation equivalents with certain syntax

- **Feature Value Augmentation**
  - Single Word Defaulting
  - Multiword Defaulting
  - Defaulting Monolingual Entries from Transfer
  - Deriving
Example for Feature Value Augmentation:

*Beitrittswerhandlung* – *accession negotiation*

- Defaulted bilingual entry:
  
  ```
  ("Beitrittswerhandlung" NST "accession negotiation" NST PRF 10000
  TAG (POL) !DATE 1296734170 !OWNER "sys" !AUTHOR "TermExtract")
  ```

- Defaulted German monolingual entry:
  
  ```
  ("Beitrittswerhandlung" NST ALO "Beitrittswerhandlung" ARGs ((N1
  (PREP "mit" "über") (CA A)) (N0 (PREP0 "über") (FCP TH) (INT T)))
  CL (P-EN S-0) GD (F) KN MS-CNT LINK S SX (N) TYN (PRO) !AUTHOR
  "TermExtract" !OWNER "sys" !DATE 1296734190)
  ```

- Defaulted English monolingual entry:
  
  ```
  ("accession negotiation" NST ALO "accession negotiation" KN CNT
  TYN (ABS PRO) MW-TYPE STRING-NST MW-BODY ((STRING
  "accession") (HEAD)) !AUTHOR "TermExtract" !OWNER "sys" !DATE
  1296734198 MW-HEAD-CAN "negotiation" MWHEAD-CAT NST)
  ```
Intelligent Terminology Extraction

Parallel Documents

Ling. Processing:
- Tokenization
- POS Tagging
- Lemmatization

Creation of RBMT Trees

Statistical Word Alignment

Initial Term Pairs

Ling. Filtering & Transformation:
- Bidirectional Reconciliation
- Collocation Check
- Ling. Driven Data Splitting

Final Term Pairs

Ling. Selection and Feature Value Augmentation

Import Files

RBMT System
## Evaluations: Extraction & Import Quantification

<table>
<thead>
<tr>
<th>Translation Direction</th>
<th>German-English</th>
<th>Spanish-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lines in Europarl Corpus</td>
<td>1,259,571</td>
<td>1,253,026</td>
</tr>
<tr>
<td>Monolingual Lists</td>
<td>441,425 / 508,592</td>
<td>406,296 / 294,069</td>
</tr>
<tr>
<td>Initial Term Pairs</td>
<td>45,857</td>
<td>35,088</td>
</tr>
<tr>
<td>Final Term Pairs</td>
<td>30,803</td>
<td>24,519</td>
</tr>
<tr>
<td>Imported Terminology</td>
<td>27,054</td>
<td>18,546</td>
</tr>
</tbody>
</table>
Translation Evaluation (1)

**Improvements:**

- Superior selection of target entities:
  - Source: prostitución infantil
  - LT-BASE: infantile prostitution
  - LiSTEX: child prostitution

- Better recognition of whole sentence structure:
  - Source: ... but I must stress that in some Member States the collection of this kind of data infringes constitutional provisions on the protection of privacy.
  - LT-BASE: ... aber ich betonen muss, dass die Sammlung dieser Art von Daten in einigen Mitgliedstaaten Verfassungsbereitstellungen auf dem Schutz von Privatleben verletzt.
  - LT-LiSTEX: ... aber ich muss betonen, dass die Sammlung dieser Art von Daten in einigen Mitgliedstaaten verfassungsrechtliche Bestimmungen auf dem Schutz der Privatsphäre verletzt.
Translation Evaluation (2)

Alternatives from TermExtraction:

- 5 best alternatives:
  - Source: ... ein massives Demokratiedefizit
  - LT-BASE: ... a massive democracy deficit
  - LT-LiSTEX: ... a massive democracy
    [deficiency|deficit|shortcoming|shortfall]

- The most frequent evaluation issue
  - Restrict number of alternatives
  - Order them by frequency
Multiword Parts Added and/or Lost:

- Alignment errors:
  
  Source: Zeitraum  
  LT-BASE: *period*  
  LT-LiSTEX: {five-year period|period}

- Free or elliptic translations in corpus:

  Source: Rechnungshof  
  LT-BASE: *court of auditors*  
  LT-LiSTEX: *Court*  

  Der Rechnungshof arbeitet jedoch mit Normen. →  
  Nonetheless, the Court is working with standards.
Multiword Expressions and Collocations:

- Large syntactic variability:
  
  Source: *des Europäischen Rates in Nizza und in Biarritz*
  
  LT-BASE: *of the European [Council|Advice] in Nice and in Biarritz*
  
  LT-LiSTEX: *of the Nice Council and in Biarritz*

- Extend intelligent representation and grammar processing for variable collocational syntax.
Wrong Capitalization of English Terms:

- Capitalized vs. non-capitalized translation alternatives:
  
  Source: ... die Verfassung
  LT-BASE: ... the [constitution|state]
  LT-LiSTEX: ... [Constitution|constitution]

➤ Only most frequent spelling variant will remain.
Translation Evaluation (6)

Wrong Translation Equivalents:

- Wrong translations in the original corpus:
  *November → October*
  *mañana → afternoon*

- Idiomatic usage in the original corpus:
  *Gleichwohl dürfte es schwierig sein, das Rad zurückzudrehen.* →
  *However, the impression is that it will be difficult to turn the clock back.*
  *Rad → clock:*
  *Der Sieg von Paolo Bettini bei der Rad-Weltmeisterschaft → The victory of Paolo Bettini during the clock world championship*

- Only very few examples
Translation Evaluation (7)

Missing Subcategorization Frames:

- Bemerkung zu jdn./etw. → remark on sb./sth.
-_extracted terminology does not include frame information.

- Differentiate merge of information from existing lexicon entries and new terms.
## Translation Quality Evaluation

<table>
<thead>
<tr>
<th></th>
<th>English ↔ Spanish</th>
<th>English ↔ German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translated TUs</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Different Translations</td>
<td>95.05%</td>
<td>95.45%</td>
</tr>
<tr>
<td>Evaluated Differences</td>
<td>287</td>
<td>970</td>
</tr>
<tr>
<td>Better</td>
<td>47.74%</td>
<td>50.82%</td>
</tr>
<tr>
<td>Equal</td>
<td>31.71%</td>
<td>41.86%</td>
</tr>
<tr>
<td>Worse</td>
<td>20.56%</td>
<td>7.32%</td>
</tr>
<tr>
<td></td>
<td>398</td>
<td>1,261</td>
</tr>
</tbody>
</table>
Conclusion and Outlook

Result:
- Great potential in multi-level hybrid approach of LiSTEX
- Viable hybrid system application
- Promising quality gains

Outlook and next steps:
- LiSTEX in other domains / corpora
- Lemmatizer and generation improvements
- Intelligent handling of syntactically free collocations
- Refine translation alternatives‘ selection
Thank you for your attention.
## Translation Quality Evaluation

<table>
<thead>
<tr>
<th>Translation Direction</th>
<th>Baseline</th>
<th></th>
<th>LiSTEX</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
</tr>
<tr>
<td>German → English</td>
<td>5.5582</td>
<td>0.1632</td>
<td>5.2430</td>
<td>0.1491</td>
</tr>
<tr>
<td>English → German</td>
<td>4.3616</td>
<td>0.1078</td>
<td>4.2718</td>
<td>0.1060</td>
</tr>
<tr>
<td>Spanish → English</td>
<td>5.8776</td>
<td>0.1953</td>
<td>5.6414</td>
<td>0.1830</td>
</tr>
<tr>
<td>English → Spanish</td>
<td>6.1375</td>
<td>0.2085</td>
<td>5.9086</td>
<td>0.1978</td>
</tr>
</tbody>
</table>