Workpackage 3
Interactive Editing

Philipp Koehn, Jesús González-Rubio

November 25, 2013
Overview

- Objective: New methods to assist the editing of translations

- Tasks in Year 2
  - Task 3.1: Sentence-level Estimate of Post-editing Work Effort (completed)
  - Task 3.2: Word-level Confidence Measures (completed)
  - Task 3.5: Display Multiple Translation Options (ongoing into year 3)
  - Task 3.6: Authoring Assistance (ongoing into year 3)
Shared Task at WMT 2013

- Much interest on confidence measures (quality estimation) in research community

⇒ Organization of shared task
  (collaboration with MateCat, MosesCore, QTLaunchPad)

- WMT workshop at ACL conference, Sofia, Bulgaria, August 2013

- Tasks
  - prediction of usefulness of sentence translations
  - prediction of post-editing time
  - word level confidence estimation

- Tasks used data from first CASMACAT field trial as test
Task 3.1

Sentence-Level Estimate of Post-Editing Work Effort
Task 3.1: Sentence-Level Estimate of Post-Editing Work Effort

- Schedule: months 1–24

- Extension of the work developed during the first year
  - Regression problem, prediction of quality scores from a set of features

- Focus: efficient management of huge sets of collinear and ambiguous features

- Proposal: A two-step training methodology
Development

- Dimensionality reduction (DR) based on *partial least squares regression* (PLSR)

- Widely-used PCA reduces dimensionality taking into account only the features
  - Reduced feature sets contain almost not redundancy...
  - but, these are not necessarily the best features to perform the prediction

- Alternatively, PLSR does take into account the values to be predicted
  - As PCA, reduced feature sets contain almost no redundancy...
  - and, the new features explain most of the variability in the values to be predicted
Contributions

- Two new DR methods based on PLSR:
  - **PLSR projection**: projects the original features into a new space (similar to PCA)
  - **Variable importance in projection**: selects a subset of the original features

- Exhaustive comparison against different widely-used DR methods
  - Measure of interest: prediction accuracy when using each DR method

- Study the influence of each DR method on the accuracy of different prediction models
• PLSR projection (PLS-P) outperformed all other tested approaches

• Huge reduction in the number of features used to perform the prediction

• Similar conclusions were obtained for all tested prediction models
Conclusions

- Projection-based DR methods usually outperformed feature selection methods
  - The proposed PLS projection outperformed widespread PCA projection

- A combination of PLSR-P and a SVM provided the best performance
  - Better prediction accuracy than models built with all the original features

- Time efficiency is a complimentary advantage of the proposed approach
  - Adequate approach to be deployed in scenarios with temporal restrictions

- The ideas explored here influenced the implementation of Active Learning in WP4
Publications


Task 3.2

Word-Level Confidence Measures
Task 3.2: Word-Level Confidence Measures

- **Schedule**: months 1–24
- Confidence estimation is addressed as a two-class classification task
- **Focus**: efficient management of large sets of noisy features
- **Proposal**: classifier based on *partial least squares discriminant analysis* (PLS-DA)
- Empirical study of the influence of context in classification accuracy
Development

• Study of new features describing context
  – Generalization of the features based on posterior probabilities
  – Context is included as an additional posterior dependency
  – Drawback: redundancy

• PLS-DA classifier
  – Performs an implicit dimensionality reduction based on PLSR projection
  – Efficient management of large sets of noisy and redundant features

• Empirical study of each individual feature

• Comparison against the state-of-the-art in word-level quality estimation
Results of the Individual Features

- Prediction accuracy degraded as context size increased
- Features based on Model 1 lexicon outperformed posterior features
Results of the PLS-DA Classifier

- PLS-DA outperformed the previously used Naïve Bayes classifier
- Robust when multiple redundant features (different context sizes) were used
Conclusions

• The best individual feature is based on a Model 1 lexicon
  – This quality estimator is thus chosen to implement Active Interaction in WP2

• Good individual performance of context-aware features, however:
  – Larger contexts reduce the individual accuracy of the features

• PLS-DA outperformed previous models
  – Better performance in all test conditions
  – Empirical robustness in the presence of redundant and noisy features
  – Scalable due to the implicit dimensionality reduction
Task 3.5

Display of Multiple Translation Options
Translation Option Array

<table>
<thead>
<tr>
<th>er</th>
<th>hat</th>
<th>seit</th>
<th>Monaten</th>
<th>geplant</th>
<th>,</th>
<th>im</th>
<th>März</th>
<th>einen</th>
<th>Vortrag</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>he has</td>
<td>for months</td>
<td>the plan</td>
<td>in March</td>
<td>a lecture</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>it has</td>
<td>for months now</td>
<td>planned</td>
<td>in March</td>
<td>a presentation</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he was</td>
<td>for several months</td>
<td>planned to</td>
<td>in the March</td>
<td>a speech</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he has made</td>
<td>since months</td>
<td>the pipeline</td>
<td>in March of</td>
<td>a statement</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he did</td>
<td>for many months</td>
<td>scheduled</td>
<td>the March</td>
<td>a general</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Main work in Year 3

- Focus
  - diversity of displayed options
  - user studies on how much to display
  - user interface issues
Translation Options in Context

Speaking in Latin to a small gathering of cardinals at the Vatican on Monday morning, Benedict said that the Church's strength lies in its ability to reflect the world as it is already capable of.

The shyness of Benedict, an often unassuming figure, is not due to weakness, but comprehensive systems - requirements, analysis and development, processing and statistical data on all chemicals, and the need for reliable and non-biased assessment and feedback, are also in place. As far as the use of information is concerned, the Church remains an essential and indispensable partner in the fight against cancer and other diseases.

An often unassuming figure, he spent much of his papacy in the shadow of his beloved predecessor.
Main Concept

- Find source word/phrase in parallel corpus

- Display possible translations
  - ranked by frequency
  - in sentence context
  - aligned phrase highlighted
Technical Aspects

- Uses parallel corpus, just as for training of machine translation systems
- Standard automatic word alignment techniques
- Data stored in suffix array

⇒ very fast retrieval
Task 3.6

Authoring Assistance
Display of Alternative Translations

Request

Norway's rakfisk: Is this the world's smelliest fish?

Norway's five million people enjoy one of the highest standards of living, not just in Europe, but in the world.

Norwegians rakfisk: - это рыбы smelliest мира?

Пять миллионов людей Норвегии наслаждаются одним из высочайших жизненных уровней не только в Европе, но в мире.

Translation matches

Norway's five million people enjoy one of the highest standards of living, not just in Europe, but in the world.

Пять миллионов людей Норвегии наслаждаются одним из высочайших жизненных уровней не только в Европе, но в мире.
Display of Alternative Translations

Display

Norway's rakfisk: Is this the world's smelliest fish?

Норвегии ракфisk: - это рыбы smelliest мира?

Пять миллионов людей Норвегии наслаждаются одним из высочайших жизненных уровней не только в Европе, но в мире.

в мире

во всем мире

в осталном мире

на всем свете

в остальных частях мира

CANCEL

ACCEPT
Relation to Paraphrasing

- Automatic paraphrasing established research topic

- Our application:
  - driven by source
  - use of search graph
  - considers sentence context
  - ranking and diversity important
  - real time performance
Generation

• User marks out part of translation (target span)

• Target span is mapped to source span

• Search graph is consulted for alternative translations for source span

• Additional translations generated by combining translation options

⇒ Initial list of translations

• Note: could also use monolingual paraphrasing resources
Components

- Partial filters: remove some translation options
- Scorers: score translations
- Filters: remove some translations from list
- Sorters: rank list
Partial Filters

Remove some bad phrase translations

- PTPF: Punctuation Partial Filter
- SDPF: String Distance Partial Filter
  remove phrase translations that are too similar to others
- FWPF: FunctionWord Partial Filter
  remove phrase translations if they have additional function words
Scorers

Scores each translation

- BFSF: Best Forward Score Function
  compares alternate translations against best path

- SDSF: Score Difference Score Function
  considers direct and indirect conditional probability

- LMSF: Language Model Score Function
Filters

Remove some translations for full span
(not just partial phrase translations)

• Span versions of phrase translation filters

• SBRF: Score Based Filter
  remove phrase translations with bad overall score
Sorters

Sorts entire final list of translations

- LMBS: Language Model Based Sorter
  uses full sentence language model score

- CBDS: Cluster Based Diversity Sorter
  first clusters translations by similarity
  then picks best translation from each cluster
  (k-means clustering)
Automatic Evaluation

• Motivation
  – alternative translations should fix translation errors
  → create bad translations by back-translation

• Process
  – Train machine translation system for both directions
  – Translate test set target → source → target*
  – Spot differences between target and target*
  – Use span in target* as “marked by user”, span in target as correct

• Experimental setting
  – WMT 2013 news translation task
  – English–Russian
  – Display 5 alternate translations (one has to be correct)
  – 2139 test cases from 1000 sentences
Example

• Translate

Unlike in Canada, the American states are responsible for the organisation of federal elections.

• Into

в отличие от канады, американские штаты ответственны за организацию федеральных выборов в соединенных штатах.

• Back into English

Unlike in Canada, US states are responsible for the organization of federal elections.
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Partial Filters</th>
<th>Score Function</th>
<th>Filters</th>
<th>Sorters</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PTPF</td>
<td>SDSF</td>
<td>-</td>
<td>-</td>
<td>135/2139</td>
</tr>
<tr>
<td>2</td>
<td>PTPF, FWPF, SDPF</td>
<td>SDSF</td>
<td>-</td>
<td>-</td>
<td>161/2139</td>
</tr>
<tr>
<td>3</td>
<td>PTPF, FWPF, SDPF</td>
<td>BFSF</td>
<td>-</td>
<td>-</td>
<td>159/2139</td>
</tr>
<tr>
<td>4</td>
<td>PTPF, FWPF, SDPF</td>
<td>LMSF</td>
<td>-</td>
<td>-</td>
<td>211/2139</td>
</tr>
<tr>
<td>5</td>
<td>PTPF, FWPF, SDPF</td>
<td>SDSF, BFSF</td>
<td>-</td>
<td>-</td>
<td>161/2139</td>
</tr>
<tr>
<td>6</td>
<td>PTPF, FWPF, SDPF</td>
<td>SDSF, LMSF, BFSF</td>
<td>-</td>
<td>-</td>
<td>218/2139</td>
</tr>
<tr>
<td>7</td>
<td>PTPF, FWPF, SDPF</td>
<td>SDSF, LMSF, BFSF</td>
<td>PTRF, FNRF</td>
<td>LMBS</td>
<td>574/2139</td>
</tr>
<tr>
<td>8</td>
<td>PTPF, FWPF, SDPF</td>
<td>SDSF, LMSF, BFSF</td>
<td>PTRF, FNRF</td>
<td>CDBS</td>
<td>493/2139</td>
</tr>
<tr>
<td>9</td>
<td>PTPF, FWPF, SDPF</td>
<td>SDSF, LMSF, BFSF</td>
<td>SBRF</td>
<td>LMBS</td>
<td>661/2139</td>
</tr>
<tr>
<td>10</td>
<td>PTPF, FWPF, SDPF</td>
<td>SDSF, LMSF, BFSF</td>
<td>SBRF</td>
<td>LMBS,CDBS</td>
<td>691/2139</td>
</tr>
</tbody>
</table>
Manual Evaluation

- Web based interactive evaluation tool

- Same setup as automatic evaluation
  - shows target span
  - 5 selectable paraphrases
  - user accepts one $\rightarrow$ correct

- Limit to three approaches

- Limit to 50 test cases
Results

- Four users (U1–U4)

- Number of instances where one translation is correct

<table>
<thead>
<tr>
<th>Method</th>
<th>U1</th>
<th>U2</th>
<th>U3</th>
<th>U4</th>
<th>average score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>6/50</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>17</td>
<td>12</td>
<td>10</td>
<td>13/50</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>20</td>
<td>26</td>
<td>29</td>
<td>26/50</td>
</tr>
</tbody>
</table>
Thank You

questions?