The INESC-ID IWSLT07 SMT System

João Graça
Diamantino Caseiro
Luísa Coheur
Outline

- INESC-ID@IWSLT
- Baseline
- Corpora
- System architecture
- Experiments
- Conclusions and future work
First Participation
- A strong motivation to build “our own” MT system
- To submerge in MT

Task
- translation of spontaneous conversations in the travel domain from Italian to English
Corpora

- **Training corpora**
  - Italian/English: 19,845 sentence pairs

- **Development corpora**
  - Dev1: IWSLT05 Written: 506 * 7
  - Dev2: IWSLT06 Speech (read): 489 sentence pairs
  - Dev3: IWSLT07 Speech (spont): 996 sentence pairs

- **Test corpora**
  - Italian/English Clean: 724 sentence pairs
  - Italian/English ASR: 724 sentence pairs
• Standard phrase-based architecture (GIZA++, Moses, SRLIM)
  – Phrase features:
    • Direct and inverse phrase probability
    • Direct and inverse IBM1 model
    • Phrase and word penalties
  – 5-gram LM
  – Minimum error training (BLEU)
  – First pass

<table>
<thead>
<tr>
<th></th>
<th>Dev1</th>
<th>Dev2</th>
<th>Dev3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>56.60</td>
<td>37.19</td>
<td>16.78</td>
</tr>
</tbody>
</table>
Outline

- INESC-ID@IWSLT
- Baseline
- Corpora
- System architecture
- Experiments
- Conclusions and future work
Outline

- INESC-ID@IWSLT
- Baseline
- Corpora
- System architecture
- Experiments
- Conclusions and future work
Experiments

1. Corpora fattening
2. Pre-processing
3. Phrase based first pass decoding
4. Filtered Phrase Table
5. Reranker
6. Post-processing
1. Corpora Fattening
1. Corpora Fattening

- Collect data in the travel domain, namely:
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
1. Corpora Fattening

• Collect data in the travel domain, namely:
  – dictionary of verb forms
    • Why?
      – As Italian is very inflected, many forms were not available on the training corpus
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
      - verbix (on-line conjugator)
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
      - verbix (on-line conjugator)
      - translation into english by off-the-shelf version of Systran
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
      - verbix (on-line conjugator)
      - translation into English by off-the-shelf version of Systran
      - manual verification
1. Corpora Fattening

Collect data in the travel domain, namely:

- dictionary of verb forms
  - Why?
    - As Italian is very inflected, many forms were not available on the training corpus
  - How?
    - Select the infinitive form of every verb in the training data
    - verbix (on-line conjugator)
    - translation into English by off-the-shelf version of Systran
    - manual verification

- a dictionary of tourism terms
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
      - verbix (on-line conjugator)
      - translation into English by off-the-shelf version of Systran
      - manual verification
  - a dictionary of tourism terms
    - Why?
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
      - verbix (on-line conjugator)
      - translation into english by off-the-shelf version of Systran
      - manual verification
  - a dictionary of tourism terms
    - Why?
      - To decrease the number of unknown nouns
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
      - verbix (on-line conjugator)
      - translation into English by off-the-shelf version of Systran
      - manual verification
  - a dictionary of tourism terms
    - Why?
      - To decrease the number of unknown nouns
    - How?
1. Corpora Fattening

- Collect data in the travel domain, namely:
  - dictionary of verb forms
    - Why?
      - As Italian is very inflected, many forms were not available on the training corpus
    - How?
      - Select the infinitive form of every verb in the training data
      - verbix (on-line conjugator)
      - translation into english by off-the-shelf version of Systran
      - manual verification
  - a dictionary of tourism terms
    - Why?
      - To decrease the number of unknown nouns
    - How?
      - Terms were collected from phrase books
1. Corpora Fattening
1. Corpora Fattening

- New data was...
1. Corpora Fattening

- New data was...
  1. used in the language model training (+data LM)
1. Corpora Fattening

- New data was...
  1. used in the language model training (+data LM)
  2. added to the corpus and used in the alignments and phrase extraction (+data Phrase)
1. Corpora Fattening

- New data was...
  1. used in the language model training (+data LM)
  2. added to the corpus and used in the alignments and phrase extraction (+data Phrase)
  3. (1) and (2) (+data Phrase and LM)
1. Corpora Fattening

- New data was...
  1. used in the language model training (+data LM)
  2. added to the corpus and used in the alignments and phrase extraction (+data Phrase)
  3. (1) and (2) (+data Phrase and LM)

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.78</td>
</tr>
<tr>
<td>+data LM</td>
<td>16.82</td>
</tr>
<tr>
<td>+data Phrase</td>
<td>16.10</td>
</tr>
<tr>
<td>+data Phrase and LM</td>
<td>16.98</td>
</tr>
</tbody>
</table>
1. Corpora Fattening

- New data was...
  1. used in the language model training (+data LM)
  2. added to the corpus and used in the alignments and phrase extraction (+data Phrase)
  3. (1) and (2) (+data Phrase and LM)

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.78</td>
</tr>
<tr>
<td>+data LM</td>
<td>16.82</td>
</tr>
<tr>
<td>+data Phrase</td>
<td>16.10</td>
</tr>
<tr>
<td>+data Phrase and LM</td>
<td>16.98</td>
</tr>
</tbody>
</table>
2. Pre-processing
2. Pre-processing

- Abbreviation expansion (as they do not appear in the speech transcription)
  - ex: Ms. --> Mister

- Some changes in the tokenization script

- Punctuation removed from Italian (source)
2. Pre-processing

- Abbreviation expansion (as they do not appear in the speech transcription)
  - ex: Ms. --> Mister

- Some changes in the tokenization script

- Punctuation removed from Italian (source)

<table>
<thead>
<tr>
<th>System</th>
<th>BLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.78</td>
</tr>
<tr>
<td>New Tokenization</td>
<td>17.22</td>
</tr>
</tbody>
</table>
2. Pre-processing

- Abbreviation expansion (as they do not appear in the speech transcription)
  - ex: Ms. --> Mister

- Some changes in the tokenization script

- Punctuation removed from Italian (source)

<table>
<thead>
<tr>
<th>System</th>
<th>BLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.78</td>
</tr>
<tr>
<td>New Tokenization</td>
<td>17.22</td>
</tr>
</tbody>
</table>
3. Phrase Based first pass decoding

- Use TreeTagger from Institute for Computational Linguistics of the University of Stuttgart (POS + lemma annotation) in 2 experiments:
  - POS distortion model
  - Lemmas for alignment
3. Phrase Based first pass decoding

- POS distortion model
3. Phrase Based first pass decoding

- POS distortion model

<table>
<thead>
<tr>
<th>Configuration</th>
<th>BLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + fat corpus - Pre-processing</td>
<td>16.98</td>
</tr>
<tr>
<td>POS distortion model</td>
<td>12.24</td>
</tr>
</tbody>
</table>
3. Phrase Based first pass decoding

- POS distortion model

<table>
<thead>
<tr>
<th>Configuration</th>
<th>BLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + fat corpus - Pre-processing</td>
<td>16.98</td>
</tr>
<tr>
<td>POS distortion model</td>
<td>12.24</td>
</tr>
</tbody>
</table>
3. Phrase Based first pass decoding
3. Phrase Based first pass decoding

- Lemmas for alignment
  - Use word lemma to improve the quality of extracted phrases (try to reduce data sparseness) both with the original corpus and with the fat corpus
3. Phrase Based first pass decoding

- Lemmas for alignment
  - Use word lemma to improve the quality of extracted phrases (try to reduce data sparseness) both with the original corpus and with the fat corpus

<table>
<thead>
<tr>
<th>Configuration</th>
<th>BLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + Fat corpus - Pre-processing</td>
<td>16.98</td>
</tr>
<tr>
<td>Baseline + Original corpus + Pre-processing</td>
<td>17.22</td>
</tr>
<tr>
<td>Lemma + Original corpus - Pre-processing</td>
<td>16.79</td>
</tr>
<tr>
<td>Lemma + Original corpus + Pre-processing</td>
<td>16.72</td>
</tr>
<tr>
<td>Lemma + Fat corpus - Pre-processing</td>
<td>16.41</td>
</tr>
<tr>
<td>Lemma + Fat corpus + Pre-processing</td>
<td>17.30</td>
</tr>
</tbody>
</table>
3. Phrase Based first pass decoding

- Lemmas for alignment
  - Use word lemma to improve the quality of extracted phrases (try to reduce data sparseness) both with the original corpus and with the fat corpus

<table>
<thead>
<tr>
<th>Configuration</th>
<th>BLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + Fat corpus - Pre-processing</td>
<td>16.98</td>
</tr>
<tr>
<td>Baseline + Original corpus + Pre-processing</td>
<td>17.22</td>
</tr>
<tr>
<td>Lemma + Original corpus - Pre-processing</td>
<td>16.79</td>
</tr>
<tr>
<td>Lemma + Original corpus + Pre-processing</td>
<td>16.72</td>
</tr>
<tr>
<td>Lemma + Fat corpus - Pre-processing</td>
<td>16.41</td>
</tr>
<tr>
<td>Lemma + Fat corpus + Pre-processing</td>
<td>17.30</td>
</tr>
</tbody>
</table>
3. Phrase Based first pass decoding

- Lemmas for alignment
  - Use word lemma to improve the quality of extracted phrases (try to reduce data sparseness) both with the original corpus and with the fat corpus

<table>
<thead>
<tr>
<th>Configuration</th>
<th>BLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + Fat corpus - Pre-processing</td>
<td>16.98</td>
</tr>
<tr>
<td><strong>Baseline + Original corpus + Pre-processing</strong></td>
<td>17.22</td>
</tr>
<tr>
<td>Lemma + Original corpus - Pre-processing</td>
<td>16.79</td>
</tr>
<tr>
<td><strong>Lemma + Original corpus + Pre-processing</strong></td>
<td>16.72</td>
</tr>
<tr>
<td>Lemma + Fat corpus - Pre-processing</td>
<td>16.41</td>
</tr>
<tr>
<td><strong>Lemma + Fat corpus + Pre-processing</strong></td>
<td>17.30</td>
</tr>
</tbody>
</table>
4. Filtered Phrase Table
4. Filtered Phrase Table

- Remove
  - all phrases with periods or question marks in the middle
4. Filtered Phrase Table

- Remove
  - all phrases with periods or question marks in the middle

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Not Filtered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + Original corpus + Preprocessing</td>
<td>17.22</td>
<td>17.45</td>
</tr>
<tr>
<td>Lemma + Original corpus + Pre-processing</td>
<td>16.72</td>
<td>16.89</td>
</tr>
<tr>
<td>Lemma + Fat corpus + Pre-processing</td>
<td>17.30</td>
<td>17.34</td>
</tr>
</tbody>
</table>
4. Filtered Phrase Table

- Remove
  - all phrases with periods or question marks in the middle

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Not Filtered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + Original corpus + Pre-processing</td>
<td>17.22</td>
<td>17.45</td>
</tr>
<tr>
<td>Lemma + Original corpus + Pre-processing</td>
<td>16.72</td>
<td>16.89</td>
</tr>
<tr>
<td>Lemma + Fat corpus + Pre-processing</td>
<td>17.30</td>
<td>17.34</td>
</tr>
</tbody>
</table>
5. Reranker

- Features according to a log-linear model in order to maximise BLEU
- 1000-best hypotheses
5. Reranker

- Sentence features:
  - first pass score
  - ratio between target and source sentence length
  - some question features
  - 3,4 and 5-grams target words LMs
  - 3,4 and 5-grams target POS LMs
  - Direct and inverse IBM1 model
  - POS similarities
5. Reranker
5. Reranker

- POS similarities
  - assume that the number of certain tags should be similar in each pair Italian/English
    - ex: NOM (it) and NNS + NN (en)
  - the Euclidean distance was used to calculate the feature score
5. Reranker

- **POS similarities**
  - assume that the number of certain tags should be similar in each pair Italian/English
    - ex: NOM (it) and NNS + NN (en)
  - the Euclidean distance was used to calculate the feature score

- **POS unlikely sequences**
  - assume that certain sequences of tags are very unlikely
    - ex: DT DT (en)
  - sentences with these sequences should be penalised
5. Reranker
5. Reranker

- Features results
  - Some features don’t give good results by its own, but are responsible for bleu increasing when combined with other features
5. Reranker

- **Features results**
  - Some features don’t give good results by its own, but are responsible for bleu increasing when combined with other features

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.45</td>
</tr>
<tr>
<td>+ length ratio</td>
<td>17.45</td>
</tr>
<tr>
<td>+ question features</td>
<td>17.51</td>
</tr>
<tr>
<td>+ word n-gram LMs</td>
<td>17.45</td>
</tr>
<tr>
<td>+ POS n-gram LMs</td>
<td>17.38</td>
</tr>
<tr>
<td>+IBM1 Dictionary</td>
<td>17.45</td>
</tr>
<tr>
<td>+POS similarity</td>
<td>17.57</td>
</tr>
<tr>
<td>All Features</td>
<td>17.66</td>
</tr>
</tbody>
</table>
5. Reranker

● Features results
  - Some features don’t give good results by its own, but are responsible for bleu increasing when combined with other features

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.45</td>
</tr>
<tr>
<td>+ length ratio</td>
<td>17.45</td>
</tr>
<tr>
<td>+ question features</td>
<td>17.51</td>
</tr>
<tr>
<td>+ word n-gram LMs</td>
<td>17.45</td>
</tr>
<tr>
<td>+ POS n-gram LMs</td>
<td>17.38</td>
</tr>
<tr>
<td>+IBM1 Dictionary</td>
<td>17.45</td>
</tr>
<tr>
<td>+POS similarity</td>
<td>17.57</td>
</tr>
<tr>
<td>All Features</td>
<td>17.66</td>
</tr>
</tbody>
</table>
6. Post-processing
6. Post-processing

- Removing leading and trailing commas
  - ex: , please ? good morning .

- Add/remove question marks or periods according with sentences types
  - ex: where ... --> where ...?

- Make changes in specific words
  - ex: good-bye --> goodbye
6. Post-processing

- Removing leading and trailing commas
  - ex: , please ? good morning .
- Add/remove question marks or periods according with sentences types
  - ex: where ... --> where ...?
- Make changes in specific words
  - ex: good-bye --> goodbye

<table>
<thead>
<tr>
<th></th>
<th>simple</th>
<th>post-processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.45</td>
<td>21.53</td>
</tr>
<tr>
<td>Reranked</td>
<td>17.55</td>
<td>21.58</td>
</tr>
</tbody>
</table>
6. Post-processing

- Removing leading and trailing commas
  - ex: , please ? good morning .

- Add/remove question marks or periods according with sentences types
  - ex: where ... --> where ...?

- Make changes in specific words
  - ex: good-bye --> goodbye

<table>
<thead>
<tr>
<th></th>
<th>simple</th>
<th>post-processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.45</td>
<td>21.53</td>
</tr>
<tr>
<td>Reranked</td>
<td>17.55</td>
<td>21.58</td>
</tr>
</tbody>
</table>
6. Post-processing

- Removing leading and trailing commas
  - ex: , please ? good morning .
- Add/remove question marks or periods according with sentences types
  - ex: where ... --> where ...?
- Make changes in specific words
  - ex: good-bye --> goodbye

<table>
<thead>
<tr>
<th></th>
<th>simple</th>
<th>post-processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.45</td>
<td>21.53</td>
</tr>
<tr>
<td>Reranked</td>
<td>17.55</td>
<td>21.58</td>
</tr>
</tbody>
</table>
Test set results
Test set results

- Primary System:
  - pre-processing + first pass + re-ranker + post-processing

- Secondary System:
  - pre-processing + first pass + post-processing
Test set results

- Primary System:
  - pre-processing + first pass + re-ranker + post-processing

- Secondary System:
  - pre-processing + first pass + post-processing

<table>
<thead>
<tr>
<th>Condition</th>
<th>Primary system</th>
<th>Secondary system</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE clean</td>
<td>26.57</td>
<td>26.35</td>
</tr>
<tr>
<td>IE ASR</td>
<td>24.16</td>
<td>24.35</td>
</tr>
</tbody>
</table>
Summary

- We introduced the INESC-ID MT system being developed at L2F (Spoken Language Systems Lab) from INESC-ID, Lisboa.
Summary

- We introduced the INESC-ID MT system being developed at L2F (Spoken Language Systems Lab) from INESC-ID, Lisboa.
- We participated in the Track of translating spontaneous conversation in the travel domain from Italian to English.
Summary

• We introduced the INESC-ID MT system being developed at L2F (Spoken Language Systems Lab) from INESC-ID, Lisboa.

• We participated in the Track of translating spontaneous conversation in the travel domain from Italian to English.

• We used a re-rank step where the 1000 n-best hypotheses were analysed. Several features were used at this step, including POS-based features.
Conclusions and Future Work
Conclusions and Future Work

Conclusions
- The re-ranker gain is not significant
- Bigger gains came from pre and post-processing of the data!!!!
Conclusions and Future Work

- Conclusions
  - The re-ranker gain is not significant
  - Bigger gains came from pre and pos- processing of the data!!!!
Conclusions and Future Work

• Conclusions
  – The re-ranker gain is not significant
  – Bigger gains came from pre and post-processing of the data!!!!

• Future Work
  – Understand what went wrong with the re-ranker
  – Perform a more systematic study of the POS-based features
  – Explore the domain adaptation
The INESC-ID IWSLT07 SMT System

João Graça
Diamantino Caseiro
Luísa Coheur