

technology  
from seed

# 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

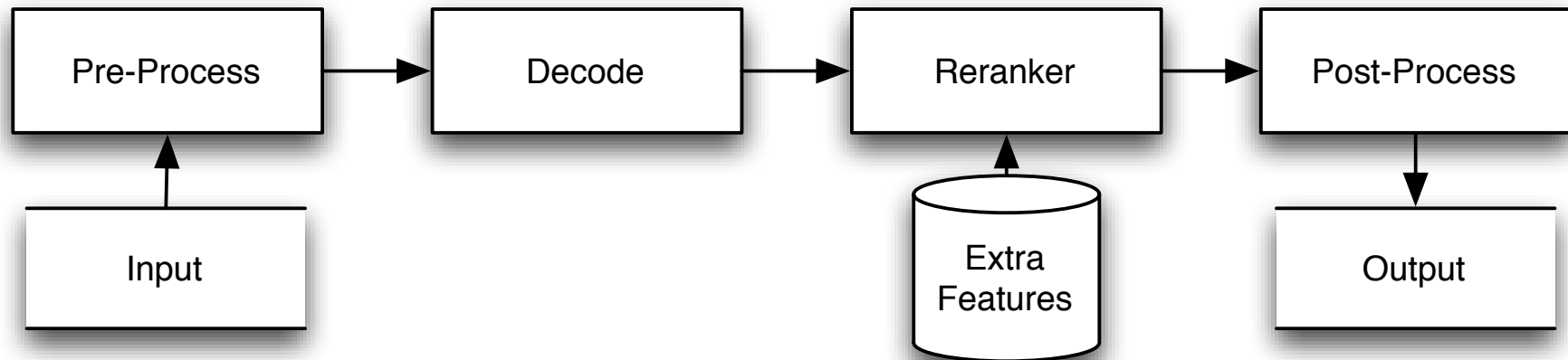
- First Participation
  - A strong motivation to build “our own” MT system
  - To submerge in MT
- Task
  - translation of spontaneous conversations in the travel domain from Italian to English

- 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

	Dev1	Dev2	Dev3
Baseline	56.60	37.19	16.78

# System architecture



# 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)

System	BLEU
Baseline	16.78
+data LM	16.82
+data Phrase	16.10
+data Phrase and LM	16.98

# 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)

System	BLEU
Baseline	16.78
+data LM	16.82
+data Phrase	16.10
+data Phrase and LM	16.98





## 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)

System	BLUE
Baseline	16.78
New Tokenization	17.22

## 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)

System	BLUE
Baseline	16.78
New Tokenization	17.22



### 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

Configuration	BLUE
Baseline + fat corpus - Pre-processing	16.98
POS distortion model	12.24

### 3. Phrase Based first pass decoding

- POS distortion model

Configuration	BLUE
Baseline + fat corpus - Pre-processing	16.98
POS distortion model	12.24



### 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

<b>Configuration</b>	<b>BLUE</b>
Baseline + Fat corpus - Pre-processing	16.98
Baseline + Original corpus + Pre-processing	17.22
Lemma + Original corpus - Pre-processing	16.79
Lemma + Original corpus + Pre-processing	16.72
Lemma + Fat corpus - Pre-processing	16.41
Lemma + Fat corpus + Pre-processing	17.30

### 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

Configuration	BLUE
Baseline + Fat corpus - Pre-processing	16.98
Baseline + Original corpus + Pre-processing	17.22
Lemma + Original corpus - Pre-processing	16.79
Lemma + Original corpus + Pre-processing	16.72
Lemma + Fat corpus - Pre-processing	16.41
Lemma + Fat corpus + Pre-processing	17.30



### 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

Configuration	BLUE
Baseline + Fat corpus - Pre-processing	16.98
<b>Baseline + Original corpus + Pre-processing</b>	17.22
Lemma + Original corpus - Pre-processing	16.79
<b>Lemma + Original corpus + Pre-processing</b>	16.72
Lemma + Fat corpus - Pre-processing	16.41
<b>Lemma + Fat corpus + Pre-processing</b>	17.30

## 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

Configuration	Not Filtered	Filtered
Baseline + Original corpus + Pre-	17.22	17.45
Lemma + Original corpus + Pre-processing	16.72	16.89
Lemma + Fat corpus + Pre-processing	17.30	17.34

## 4. Filtered Phrase Table

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

Configuration	Not Filtered	Filtered
Baseline + Original corpus + Pre-	17.22	17.45
Lemma + Original corpus + Pre-processing	16.72	16.89
Lemma + Fat corpus + Pre-processing	17.30	17.34





## 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

System	BLEU
Baseline	17.45
+ length ratio	17.45
+ question features	17.51
+ word n-gram LMs	17.45
+ POS n-gram LMs	17.38
+IBM1 Dictionary	17.45
+POS similarity	17.57
All Features	17.66



## 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

System	BLEU
Baseline	17.45
+ length ratio	17.45
+ question features	17.51
+ word n-gram LMs	17.45
+ POS n-gram LMs	17.38
+IBM1 Dictionary	17.45
+POS similarity	17.57
All Features	17.66



## 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

	simple	post-processed
Baseline	17.45	21.53
Reranked	17.55	21.58

## 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

	simple	post-processed
Baseline	17.45	21.53
Reranked	17.55	21.58



## 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

	simple	post-processed
Baseline	17.45	21.53
Reranked	17.55	21.58



# 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

Condition	Primary system	Secondary system
IE clean	26.57	26.35
IE ASR	24.16	24.35

# Summary



## 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 pos- 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 pos- 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*

