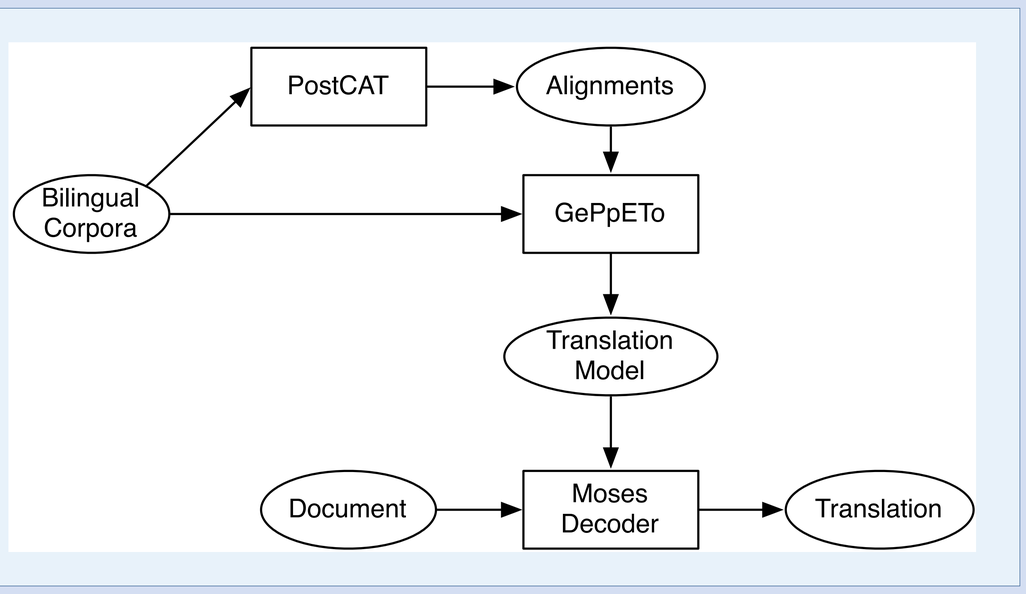
The INESC-ID Machine Translation System for the IWSLT 2010

Wang Ling, L²F, INESC-ID/IST
Tiago Luís, L²F, INESC-ID/IST
João Graça, L²F, INESC-ID/IST
Luísa Coheur, L²F, INESC-ID/IST
Isabel Trancoso, L²F, INESC-ID/IST

Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa
Laboratório de Sistemas de Língua Falada

{wang.ling,tiago.luis,joao.graca,luisa.coheur,isabel.trancoso}@l2f.inesc-id.pt

Overview



- PostCAT - Constrained Alignment Toolkit
- Geppetto - General Phrase Extraction Toolkit

Geppetto

Phrase extraction toolkit with the following extensions:

- Weighted Alignments - Extracts phrase pairs using posterior alignments
- Punctuation Filtering - Filters phrases based on punctuation

1

Constrained Alignments

PostCAT

- Uses a Hidden Markov Model for estimating alignments
- Produces posterior probabilities over alignments
- Allows the definition of constraints for the posteriors

Constrains

- Regular HMM (HMM) - Regular HMM model with no constraints
- Bijective Constrains HMM (BHMM) - Gives higher posterior probabilities to alignments that use large numbers of source words
- Symmetric Constrains HMM (SHMM) - Gives higher posterior probabilities to alignments that also aligned in the reverse model

2

Weighted Alignments & Punctuation Filtering

Geppetto

- Phrase extraction toolkit with an easily extensible interface
- Defines key control points that can manipulate the behavior of the phrase extraction and translation model creation
- Default behavior is identical to the training scripts in Moses

Weighted Alignment Matrices

- Use posterior probabilities of the alignments instead of the best alignment
- Posterior probabilities are retrieved by the PostCAT toolkit

Punctuation Filtering

- no-terminal-punct - No terminal punctuation (".", "!" and "?") is accepted
- no-terminal-punct-unless-last - No terminal punctuation is accepted unless it is the last token

3

Experimental Results

Constrained Alignments

Alignment	BTEC (Fr-En)	DIALOG (Cn-En)	DIALOG (En-Cn)
HMM	59.93	38.49	**45.47**
BHMM	62.45	38.17	44.43
SHMM	**62.46**	**41.42**	44.99

- Posterior threshold set to 0.4
- In HMM the EM algorithm was run for 5 iterations
- In the BHMM and SHMM the EM algorithm was run for 2 iterations
- In DIALOG (En-Cn) task the results were higher for the HMM due to the lack of stability of the tuning algorithm (after using stabilization algorithms the results became consistent with the other tasks)

Weighted Alignment Matrices

Alignment	BTEC (Fr-En)	DIALOG (Cn-En)	DIALOG (En-Cn)
HMM-post	61.74	39.48	**46.11**
BHMM-post	62.74	40.69	45.23
SHMM-post	**63.07**	**42.15**	45.00

- Phrase acceptance threshold was set to 0.1

Punctuation Filtering

Acceptor	BTEC (Fr-En)	DIALOG (Cn-En)	DIALOG (En-Cn)
base (SHMM-post)	63.07	**42.15**	45.00
no-terminal-punct	**63.41**	41.44	46.68
no-terminal-punct-unless-last	62.62	41.24	**46.86**

- Punctuation filtering improved results in BTEC (Fr-En) and DIALOG (En-Cn)

4

Official Results

Primary Submission

no_case+no_punc 1best	BTEC (Fr-En)	DIALOG (Cn-En)	DIALOG (En-Cn)
IWSLT 2010	50.29	18.03	24.69
IWSLT 2009	57.50	28.73	34.28

- Setup - Symmetric constrained alignments, weighted posteriors and punctuation filtering

5

Conclusions

System focused on refining the Translation Model using:

- Constrained Alignments
- Weighted Alignment Matrices
- Punctuation Filtering

Used toolkits are available at:

- PostCAT - *http://www.seas.upenn.edu/~strctlrn/CAT/CAT.html*
- Geppetto - *http://code.google.com/p/geppetto*

6

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007. The PhD thesis of Tiago Luís is supported by FCT grant SFRH/BD/62151/2009. The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010.