Karlsruhe Institute of Technology

KIT EN-FR Systems for the IWSLT 2012

Mohammed Mediani, Yuqi Zhang, Thanh-Le Ha, Jan Niehues, Eunah Cho, Teresa Herrmann, Rainer Kärgel and Alex Waibel

Outline

- System summary
- Preprocessing
- POS-based reordering
- Adaptation
- Additional components
  - Bilingual language models (BiLM)
  - Cluster LM
  - Discriminative word lexicon (DWL)
  - Continuous space language models (RBMLM)
- Postprocessing
- Experiments and Results
- Appendix

System summary

- Phrase-based system trained on:
  - TED, EPPS, NC, and Giga
- Modified Kneser-Ney smoothed probabilities (Translation)
- 4-gram language models trained on:
  - Parallel data, News shuffled
- Tuned towards Dev 2010 using MERT
- POS-based reordering
- Adaptation
- Additional models:
  - BiLM
  - Cluster LM
  - RBMLM
  - DWL
- Postprocessing

Newly introduced models in IWSLT2012

- Union Candidate Selection translation model adaptation (CSUnion)
- Continuous space language models (RBMLM)
- Postprocessing (POS-based agreement correction)

Preprocessing

- Remove long sentences
- Remove sentences with length mismatch
- Filter the Giga corpus:
  - Train an SVM classifier to filter out non-parallel pairs
- Parallel data:

Corpora	Original (x10^6)			Training (x10^6)		
	#Pairs	#EN words	#FR words	#Pairs	#EN words	#FR words
TED	0.15	2.41	2.48	0.14	2.80	2.96
EPPS	2.00	50.20	51.39	1.98	54.57	58.93
NC	0.14	2.99	3.37	0.14	3.44	3.93
Giga	22.52	572.40	653.36	16.80	446.90	516.56

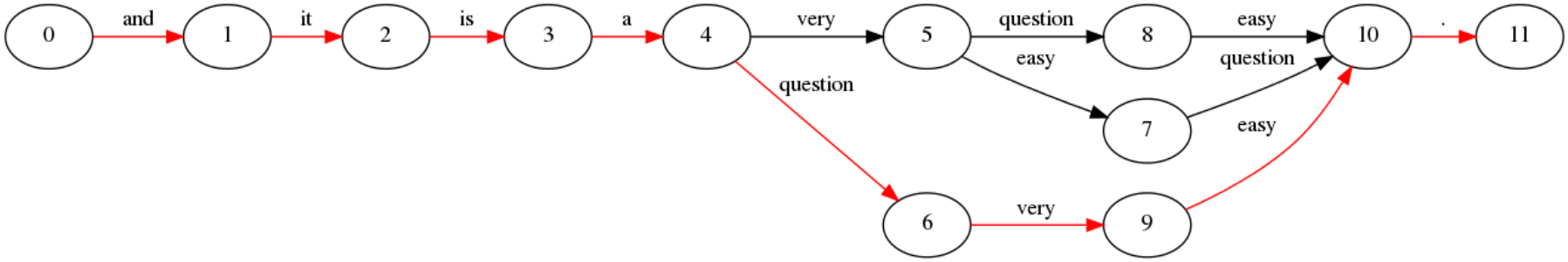
Preprocessing

- Remove long sentences
- Remove sentences with length mismatch
- Filter the Giga corpus:
  - Train an SVM classifier to filter out non-parallel pairs
- Parallel data:

- For SLT system:
  - Lowercase the source side
  - Remove all source punctuations except period

POS-based reordering

- Rules to reorder source side are learnt based on POS tags
- Rules are applied to source side
- Best reordering alternatives are recorded in a lattice

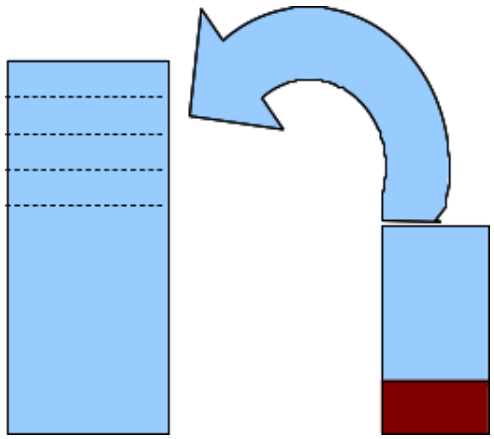


Adaptation

- LM Adaptation:
  - In-domain language model (Trained on TED) as extra model
  - Weights are tuned log-linearly

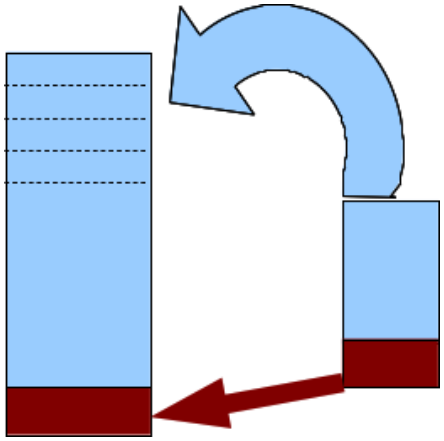
Adaptation

- LM Adaptation:
  - In-domain language model (Trained on TED) as extra model
  - Weights are tuned log-linearly
- TM adaptation
  - Extend the out-of-domain TM with scores from the in-domain model



Adaptation

- LM Adaptation:
  - In-domain language model (Trained on TED) as extra model
  - Weights are tuned log-linearly
- TM adaptation
  - Extend the out-of-domain TM with scores from the in-domain model
- Union Candidate Selection
  - Take the union of phrase pairs from in-domain and out-of-domain models



Additional models

- **Bilingual language model**
  - Wider context for the decoder
  - A language model containing target words together with their aligned source words
  - Introduced as an additional factor in the translation model

```
, ancient # || ancien | ancien_ancient ||

, ancient # || anciennes | anciennes_ancient ||

, and , instead of # || , et , au de | ,_, et_and ,_, au_instead lieu_instead
de_of ||

, although that # || , bien qu' | ,_, bien_although  qu'_that ||
```

Additional models

- **Cluster language model**
  - Language model containing the classes of the target words
  - Classes are generated using MKCLS algorithm
  - Classes trained on TED only

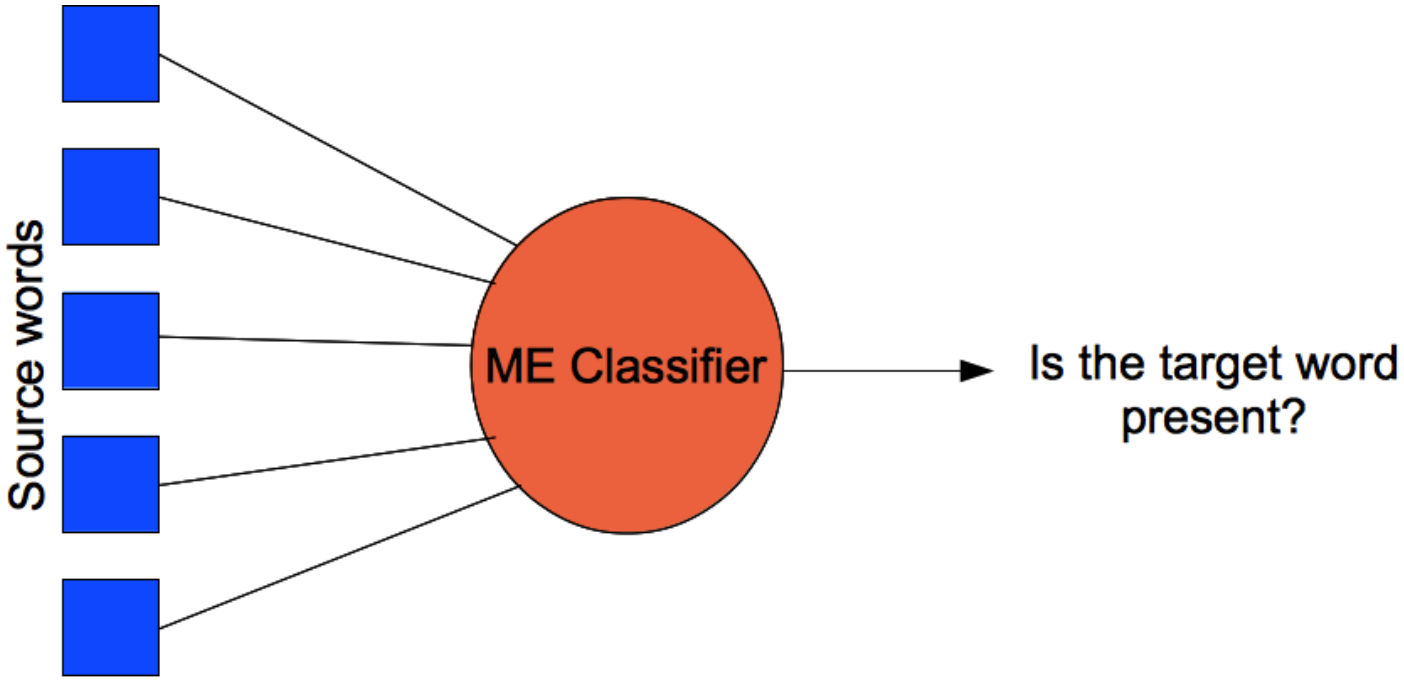
```
, health care , # || , | 19 || aux | 2 || soins | 51 || de | 20 || santé | 44 || ,
| 19 ||

, he went # || , | 19 || est | 28 || allé | 32 ||

, not because I am # || , | 19 || non | 13 || pas | 42 || que | 14 || je | 3 ||
sois | 28 ||
```

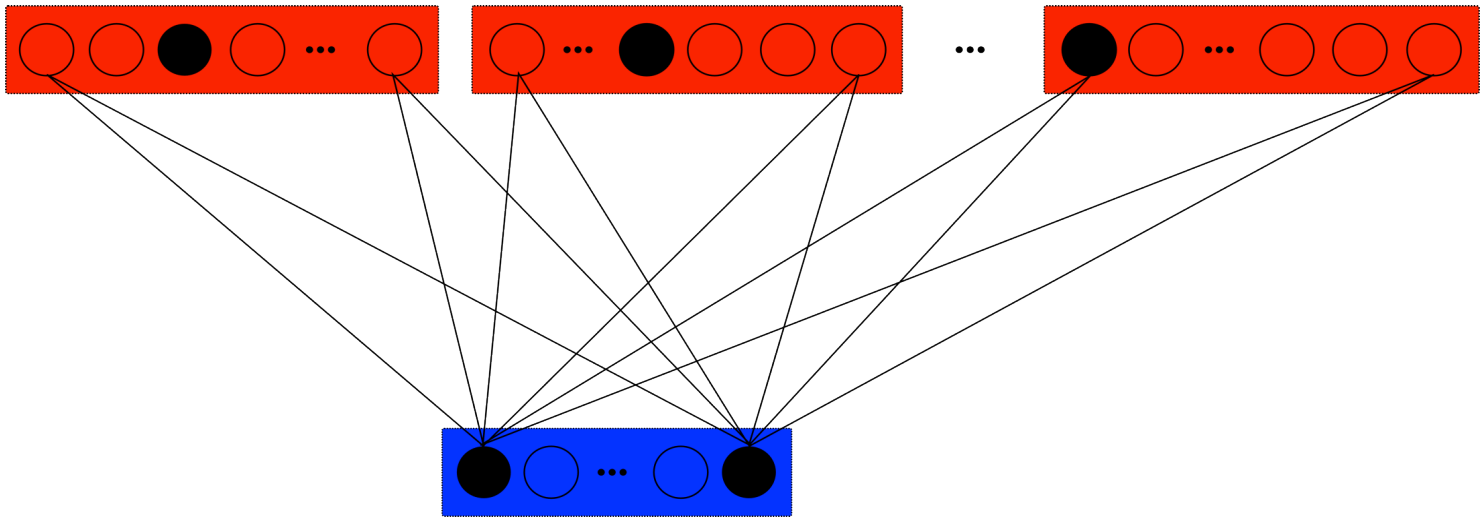
Additional models

- **Discriminative word lexicon**
  - Maximum-entropy classifier for each target word
  - Source words are the features
  - Trained on TED only



Additional models

- **Continuous space language model**
  - A restricted Boltzman machine neural network LM
  - A context of 8 words
  - Trained on TED only



Postprocessing

- Restricted heuristics for **agreement correction**
  - Based on POS-tags of the generated hypothesis
  - Corrections are set in accordance to a noun applied on its surrounding words
  - Correct adjectives in case: ADJ NOUN or NOUN ADJ
  - Correct articles, possessive, and *quelque*, if it is immediately before or an adjective in between
  - Correct past participle in case: NOUN être PP
  - Examples in Appendix

Experiments and Results

- MT system

System	Dev 2010	Test 2010
Baseline	28.50	31.73
+Bilingual LM	28.93	31.90
+Cluster LM	29.15	32.13
+CSUnion	29.27	32.21
+DWL	29.37	32.70
+RBM LM	29.46	32.78
+Agreement Correction	-	**32.84**

- UN data and Google n-grams were not helpful

Results

- SLT system

			Optimization on text		Optimization on ASR
System	Dev2010 (Text)	Test2010 (Text)	Test2010 (ASR)	Dev2010 (ASR)	Test2010 (ASR)
Baseline	25.37	27.57	21.68	19.11	21.86
+Adaptation	25.64	28.08	21.90	19.31	22.04
+Bilingual LM	25.07	28.08	22.07	19.14	22.28
+Cluster LM	25.17	28.79	22.57	19.32	22.40
+DWL	25.06	28.84	22.79	19.34	22.23
+Agr. Correction	-	-	**22.86**	-	-

- The Giga data was not helpful in this system

Appendix

SVM filtering

- Given a pair, select one of two classes Reject=-1, Keep=1
- Features considered:
  - Difference in number of words between source and target
  - IBM 1 score (both direction)
  - #unaligned words (source and target)
  - Maximum fertility (source and target)

Precision (%)	Recall (%)	F-Score (%)
98.45	92.00	95.12

The effect of the Giga corpus

- In the MT system

System	Dev2010	Test2010
Baseline	28.29	31.11
+Giga	28.50	31.73

- In the SLT system

System	Dev2010	Test2010
Baseline	18.93	21.84
+Giga	18.67	21.08

Translation examples

- CS Union Examples

WITHOUT CSUNION: Ce sont des patients subissant une procédure douloureuse .

WITH CSUNION: Ce sont des vrais patients subissant une procédure douloureuse .

REF: Des patients réels subissent une opération douloureuse.

WITHOUT CSUNION: Il y a des records sur ce point .

WITH CSUNION: Il y a des dossiers mondiaux sur ce point .

REF: Il y a aussi les records du monde.

Translation examples

- Continuous Space LM example

WITHOUT RBMLM: J'en ai compté le nombre de livres avec " bonheur " dans le titre publié dans les cinq dernières années et ils ont abandonné après environ 40 , et il y avait beaucoup d'autres .

WITH RBMLM: J'en ai fait compter le nombre de livres avec " bonheur " dans le titre publié au cours des cinq dernières années et ils ont abandonné après environ 40 , et il y avait beaucoup plus .

REF: Il y a quelqu'un qui voulait compter le nombre de livres publiés au cours des 5 dernières années, dont le titre contenait "bonheur". Il a abandonné au bout du 40ème, il y en avait bien plus.

Translation examples

- Agreement correction examples

HYP: Une capsule coloré , c'est jaune d'un côté et rouge sur l'autre est mieux qu'une capsule blanc .

AGR. CORRECT: Une capsule colorée , c'est jaune d'un côté et rouge sur l'autre est mieux qu'une capsule blanche .

HYP: Imaginez que votre prochaine vacances vous savez qu'à la fin des vacances tous vos images sera détruite , et vous obtiendrez un médicament ingestion de sorte que vous ne me rappelle pas rien .

AGR. CORRECT: Imaginez que vos prochaines vacances vous savez qu'à la fin des vacances tous vos images sera détruite , et vous obtiendrez un médicament ingestion de sorte que vous ne me rappelle pas rien .