The DCU Machine Translation Systems for IWSLT 2010

Hala Almaghout Jie Jiang Andy Way

CNGL, School of Computing, Dublin City University

{halmaghout, jjiang, away}@computing.dcu.ie

We give a description of the DCU machine translation systems submitted to the evaluation campaign of The International Workshop on Spoken Language Translation (IWSLT) 2010. We participated in the BTEC Arabic-to-English task in addition to the DIALOG task for translation between English and Chinese in both directions. We explore different extensions to PB and HPB Machine Translation Systems. We deploy a paraphrase system as an extension to our English-to-Chinese PB translation system. For the HPB system, two different syntactic augmentation methods are investigated: the first is Syntax-Augmented Machine Translation, which uses syntactic labels extracted from constituent grammar, while the other one uses syntactic labels extracted from Combinatory Categorial Grammar. In addition, we combine the output of our hierarchical systems using a system combination method based on confusion networks.

Translation Tasks

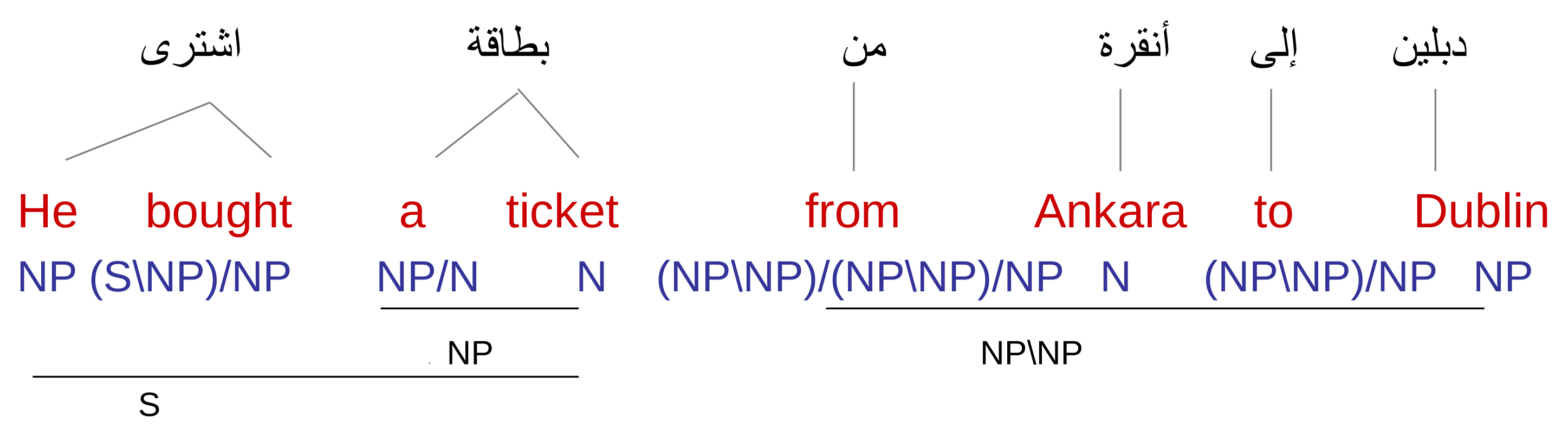
- **BTEC Task Arabic-English**
- **DIALOG Task Chinese ↔ English**

Data Pre-processing

- **Arabic Data**
  - Morphological segmentation of clitics.
  - Several segmentation schemes were examined on IWSLT data. We chose MADA D3 segmentation scheme which achieved the best BLEU score.
- **Chinese Data**
  - Word re-segmentation using the ICTCLAS tool2 .
  - Heuristic rules were used to adjust the segmentation results for Chinese numbers.
  - Conversion of punctuation marks, numbers and Latin letters from Chinese form into Latin form.
- **Case and Punctuation Restoration**
  - **Case Restoration:** using phrase-based translation model.
  - **Punctuation Restoration**: hidden n-gram tool from SRILM

CCG-Augmented Machine Translation System

- CCG-based syntactic labels extracted from CCG super tagged target side are attached to non-terminal in hierarchical rules.
- Using CCG flexible structures and rich syntactic supertags our system is able to label more phrases with CCG-based labels that reflect correctly the syntactic constraints imposed on phrases.


اشترى
بطاقة
من
أنقرة
إلى
دبلين
He bought a ticket from Ankara to Dublin
NP (S\NP)/NP NP/N N (NP\NP)/(NP\NP)/NP N (NP\NP)/NP NP
NP
NP\NP
S

S→ (He bought a ticket , اشترى بطاقة)
NP\NP→ (from Ankara to Dublin ,من أنقرة إلى دبلين)
NP\NP→(from NP to NP , منNP إلى NP)
S→ (He bought NP, اشترىNP)
S→ (He bought NP from NP to NP, اشترىNP من NP إلى NP)

Translation Systems

- **Hierarchical Phrase-Based System**
- **Syntax-Augmented Machine Translation System (SAMT)**
- **CCG-Augmented Machine Translation System**
- **System Combination:** is built using the MANY tool which is based on decoding a lattice made of several confusion networks. MANY parameters are optimized using CONDOR.

- **Paraphrase MT system for English-Chinese:**
  - Source-language paraphrases are used.
  - Paraphrase options are represented as source-side lattices
  - Probabilities on lattice edges are formed to penalize any path going through the paraphrase option.

Experiments Results

System	CE (CRR)	CE (ASR)	AE
HPB	13.58	12.79	**46.25**
CCG	**14.21**	**12.96**	45.30
SAMT4	13.75	12.86	46.06
Syscomb	13.96	12.69	46.11

DIALOG Task Chinese-English

- CCG-based system achieved the best performance beating HPB by 4.63%, 1.32% relative BLEU on CRR and ASR tasks, respectively.
- No improvement for system combination

BTEC Task Arabic-English

- HPB system achieved the best performance
- This might be due to smaller AE training data in comparison with CE.
- No improvement for system combination

DIALOG Task English-Chinese

- Paraphrase system achieved better performance than PB system by 4.42%, 6.35% relative BLEU on CRR and ASR task, respectively.

System	EC(CRR)	EC(ASR)
PB	20.13	17.32
Paraphrase	**21.02**	**18.42**

Conclusion

- Constituent-based and CCG-based syntactic extensions to HPB in addition to PB paraphrase system were examined.
- CCG-based HPB achieved better performance than HPB on CE translation under both ASR and CRR conditions.
- Paraphrase system outperformed PB system on EC translation under both ASR and CRR conditions.
- No performance gain was achieved for system combination.
- Having no improvement of syntactic extension to HPB on AE task might be due to the fact the AE training data has about third of the sentences of CE training data.

This research is supported by the Science Foundation Ireland (Grant 77/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

This research is funded by Science Foundation Ireland (SFI)
as part of the National Development Plan (NDP) 2007-2013