

DCU MT System and Recent Research Improvements

Jinhua Du, Yifan He, Ankit K. Srivastava, Rejwanul Haque,
Sandipan Dandapat, Andy Way

CnGL, School of Computing, Dublin City University, Dublin, Ireland

1. Introduction

- **MaTrEx**: multi-engine MT system developed at DCU, which exploits PB-SMT, Augmented PB-SMT, Hiero, EBMT and system combination techniques to build a cascaded framework.
- **Research**: we exploited some new techniques in our system building such as source-side context-informed MT, Syntax-augmented MT and METEOR-based MERT which significantly augmented our system.

2. System Architecture [1]

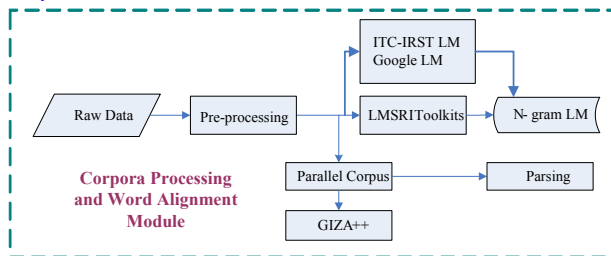


Figure 1: Data Pre-processing & Word Alignment

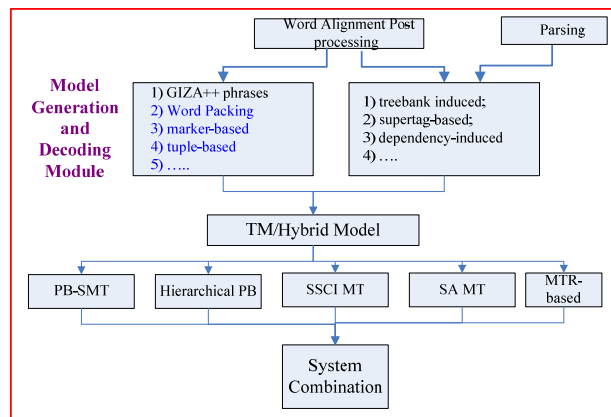
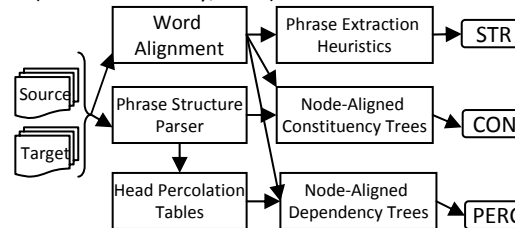


Figure 2: Model Building & Translation Process

3. Syntax-Augmented (SA) MT [2]

We augment the phrase-based translation model with syntactically motivated phrase pairs extracted from node-aligned parallel treebanks (Srivastava & Way, 2009).



➤ Obtain 3 types of phrase tables: STR, CON, PERC

➤ Merge the 3 tables by concatenating phrase pairs and re-estimating probabilities.

➤ All other functioning is similar to that in baseline PB-SMT system.

5. METEOR-based MERT [4]

In addition to tuning on BLEU, we also tune on METEOR to obtain more diverse translation outputs. We use a larger chunk penalty than the default parameters of METEOR to avoid unjustly verbose translations, in:

$$Meteor = F \cdot (1 - \gamma \cdot (\frac{\#chunks}{\#matches})^\beta), \text{ we set } \gamma=1 \text{ (0.28 in default)}$$

As shown below, the output is still longer than the reference, which results in high METEOR (good) and TER (not good).

4. Source-side context-informed (SSCI) MT [3]

➤ Context Information

$$\text{Lexical } CI = \{f_{i_k-1}, \dots, f_{i_k-1}, \hat{f}_k, f_{j_k+1}, \dots, f_{j_k+1}\}$$

$$\text{Syntactic } CI = \{SI(f_{i_k-1}), \dots, SI(f_{i_k-1}), \hat{f}_k, SI(f_{j_k+1}), \dots, SI(f_{j_k+1})\}$$

➤ Context Informed Features

$$h_{mbf} = \log P(\hat{e}_k | \hat{f}_k, CI(\hat{f}_k))$$

Context-informed features are expressed as the conditional probability of the target phrase \hat{e}_k given the source phrase \hat{f}_k and its context information $CI(\hat{f}_k)$. To avoid sparseness problems, the probability is estimated using **TIMBL** which includes three different classifiers: **IGTREE**, **IB1** and **TRIBL** (Daelemans, 2005).

Derived Memory-based feature h_{mbf} is directly integrated into **Moses**.

Official results in NIST 2009 Evaluation: (DCU System)

	BLEU	TER	METEOR
BLEU-tuned	0.2205	0.6568	0.4763
METEOR-tuned	0.2169	0.7562	0.5024

Reference

[1] Du, J., Y. He, S. Penkale and A. Way. 2009. MaTrEx: the DCU MT System for WMT 2009. In Proceedings of the Third Workshop on Statistical Machine Translation, EACL 2009, Athens, Greece. pp 95-99.

[2] Srivastava, A. and A. Way. 2009. Using Percolated Dependencies for Phrase Extraction in SMT. In Proceedings of the Twelfth Machine Translation Summit (MT Summit 2009), Ottawa, Canada.

[3] Haque, R., S. Naskar, Y. Ma and A. Way. 2009. Using Supertags as Source Language Context in SMT. In Proceedings of the 13th Annual Meeting of the EAMT-2009, Barcelona, Spain. pp 234-241.

[4] He, Y. and A. Way. 2009. Improving the Objective Function in Minimum Error Rate Training. In Proceedings of the Twelfth Machine Translation Summit (MT Summit 2009), Ottawa, Canada.