Exploiting Alignment Techniques in MaTrEx: the DCU MT System for IWSLT 2008

Yanjun Ma†, John Tinsley†, Hany Hassan†, Jinhua Du‡, Andy Way†‡
National Centre for Language Technology† and Centre for Next Generation Localisation‡
School of Computing,
Dublin City University, Dublin 9, Ireland
{jtinsley, yma, hhassan, jdu, away}@computing.dcu.ie

Introduction

MaTrEx (Machine Translation using Examples) is a hybrid system which can exploit EBMT, SMT and syntax-based techniques to build a combined translation model. MaTrEx is built following established Design Patterns and consists of a number of extensible and re-implementable modules. Some significant modules include:

- **Word Alignment Module:** outputs a set of word alignments given a parallel corpus;
- **Chunking Module:** outputs a set of chunks given an input corpus;
- **Chunk Alignment Module:** outputs aligned chunk pairs given source and target chunks from comparable corpora;
- **Decoder:** returns optimal translation given a set of aligned sentences, chunk/phrase and word pairs.

New Alignment Techniques

Word Packing

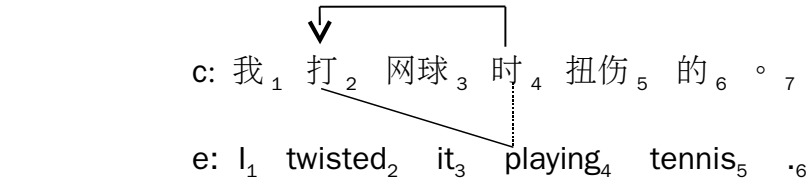
- **Candidate Extraction**

白葡萄酒：white wine
抱歉：excuse me
报警：call the police
fifteen: 十 五
here: 在 这里

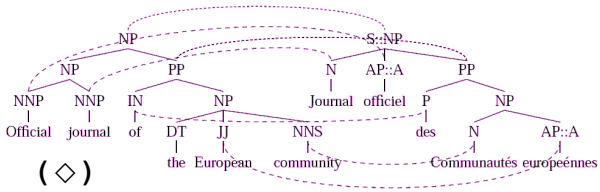
- **Reliability Estimation**
- **Bootstrapping Word Alignment**

Syntax-enhanced Word Alignment

- **Anchor Word Alignment**
- **Discriminative Syntax-Enhanced Word Alignment**
- **Search**

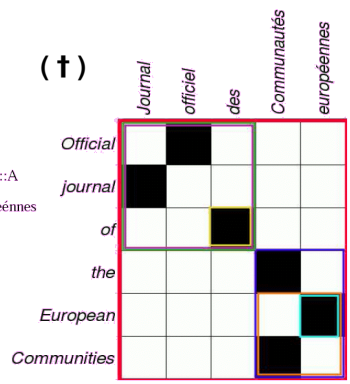


Treebank-based Phrase Alignment



Punctuation Restoration

- **Translation-based Punctuation Restoration**
- **Majority voting techniques to restore the final punctuation mark**



† Official journal ↔ Journal officiel
† Official journal of ↔ Journal officiel des
∗ Official journal of the/ European Communities ↔ Journal officiel des/ Communautés européennes
∗ of ↔ des
∗ of the European Communities ↔ des Communautés européennes
∗ the European Communities ↔ Communautés européennes
∗ European ↔ européennes
◊ Communities ↔ Communautés
◊ Official ↔ officiel
◊ journal ↔ Journal

(* = extracted from both † and ◇)

Official Results

System	Challenge Task				BTEC						Pivot	
	ZH-EN		EN-ZH		AR-EN		ZH-EN		ZH-ES		ZH-EN-ES	
	CRR	ASR-1	CRR	ASR-1	CRR	ASR-1	CRR	ASR-1	CRR	ASR-1	CRR	ASR-1
Baseline	31.94	27.14	40.80	34.29	–	–	35.95	31.45	26.93	–	28.32	–
Word Packing	29.67	26.76	40.04	34.33	–	–	35.22	31.51	–	–	–	–
Syntax-Enhanced	34.52	29.31	42.43		–	–	38.23	34.23	–	–	–	–
Treebank	28.81	26.53	37.73	32.82	–	–	37.85	32.42	29.24	26.70	32.92	29.48
OOV Smoothing	32.59	–	–	–	–	–	–	–	–	–	–	–
All Smoothing	23.95	–	–	–	47.15	38.58	–	–	–	–	–	–
Data Combo	36.40	30.86	46.30	40.22	–	–	39.66	33.97	–	–	–	–

Discussion

- Proper training on large data is always preferential.
- Syntax-enhanced word alignment leads to improvements.
- Treebank phrase extraction improves inconsistently
- Word packing leads to drops in performance.
- Smoothing and case/punctuation restoration techniques also effective.

science foundation ireland