Auto Word Alignment Based Chinese-English EBMT

Yang Muyun, Zhao Tiejun et al

Machine Intelligence & Translation Lab

Research Center for Language Technology

School of Computer Science & Technology

Harbin Institute of Technology,China

Contents

- Background
  - Introduction to MI&T Lab
  - Recent work
- Word Alignment Based EBMT
- Experiments and Discussions
- Conclusion

Background

- About MI&T LAB
  - Began Chinese-English MT in 1987
  - First CEMT system of the mainland in 1988
  - Top CEMT in MT Evaluation 1996 Held by National 863 Project
  - Joint MT Lab with MSRC in 2000
  - Joint NLP&Speech Lab with MSRA in 2004

Background

- MT in the MS-HIT Joint Lab
  - Conquer Barrier between Chinese English by Bilingual Corpus Based Knowledge Acquisition (KA)
  - From covert sentence pairs to overt translation knowledge
    - Least knowledge required→ Statistical MT(or TM)
    - Some knowledge required→ EBMT
    - Intensive knowledge required→ RBMT

Background

- MT in the MS-HIT Joint Lab
  - 2000-2001: Chinese English Bilingual Corpus Processing
    - Dictionary based sentence alignment;
    - Hybrid strategy for word alignment;
    - 100,000 Chinese English sentence beads;
    - Word aligned Chinese English corpus (60,000 beads)

Background

- MT in the MS-HIT Joint Lab
  - 2001-2002: Auto KA Based MT
    - Mono-lingual parsing based Structure Alignment(1 Coling'02 paper)
    - Auto template acquisition based ECMT
    - MT Evaluation Methods (1 Coling'02 paper)
  - 2002-2003: Chinese parsing
    - Chinese treebank (30,000 with Base Phrases,Head)
    - Head-driven Model for Chinese Parsing
    - Word alignment based EBMT

Contents

- Background
- Word Alignment Based EBMT
  - Introduction
  - Word Alignment Based Example Extraction
  - Finding Right Examples
  - Translation Selection
- Experiments and Discussions
- Conclusion

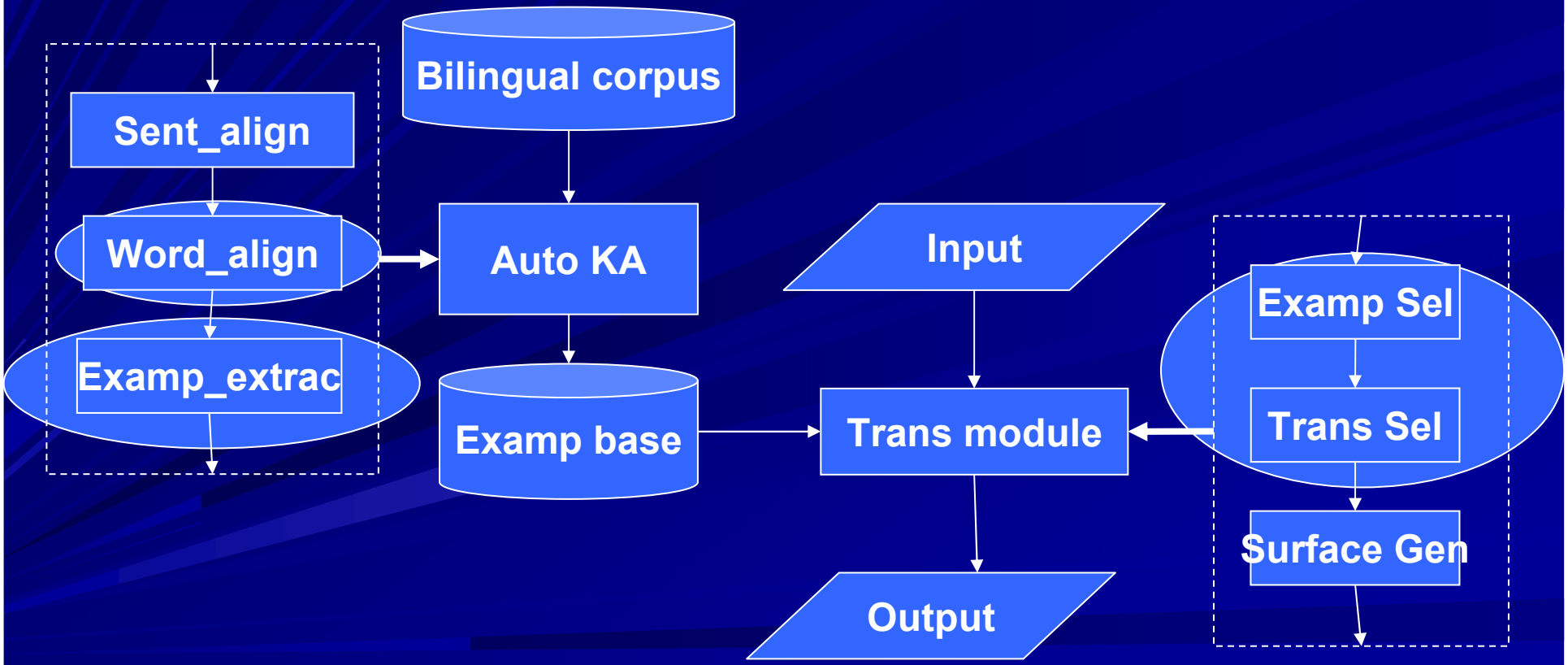
Word Alignment Based EBMT

- Introduction
  - Auto construction: least manual work;
  - Sub-sentential focus: phrase level example;
  - Adaptability: domain, (language if possible);
  - Linguistic light approach: less information loss;

Word Alignment Based EBMT

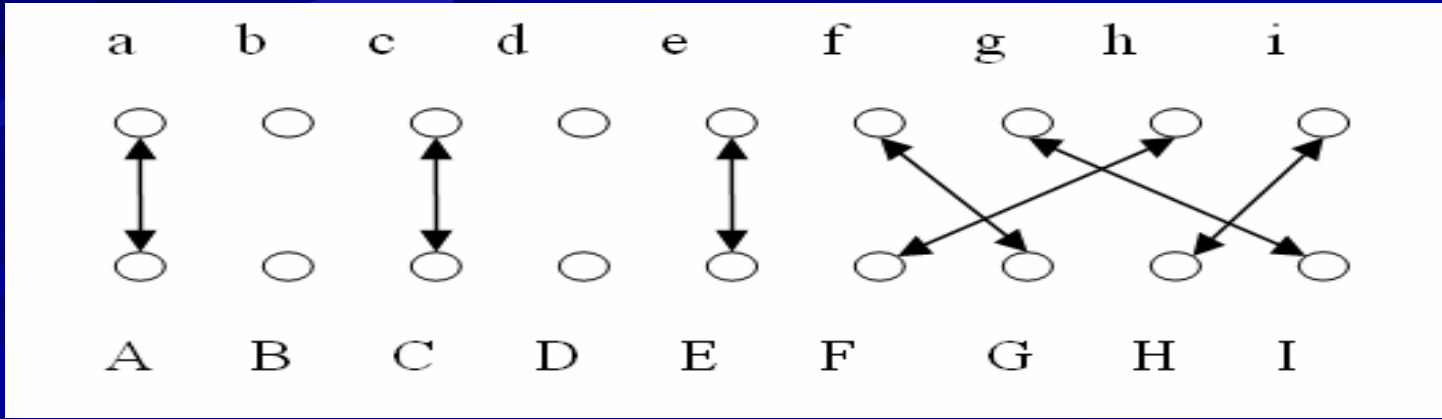
- EBMT vs Segmentation (Dic ? Example_base)
  - Input: 您的登山小组有几个人？
  - Word_Seg: 您/的/登/山/小组/有/几/个/人/？
  - Example_Seg: 您的/ 登山小组/ 有几个人/ ？
  - Translation: your/ climbing group/ how many people are there/ ?
  - Final: How many people are there in your climbing group ?

Word Alignment Based EBMT



Word Alignment Based EBMT

- Word alignment based example extraction
  - **Atomic** (aligned words)**:** (a-A)(c-C) (e-E) (f-G) (g-I) (h-F) (i- H)
  - **Parallel extension:** (ab-AB) (bc-BC) (bcd-BCD) (cd-CD) (de-DE)
  - **Locked/non-parallel:** (fghi-FGHI)



Word Alignment Based EBMT

- Finding right examples
  - Example length: bigger context;
  - Segment (concatenated examples from same sentence) length: consistency;
  - Word links: better translation correspondence;
  - Frequency: statistically reliable;

Word Alignment Based EBMT

- Finding right examples

$$\overline{Segment} = \arg\max_{\substack{0<l<n+1 \\ k_{i-1}<k_i}} \sum_{i=0}^{l} \delta([s_{k_{i-1}+1} \ldots s_{k_i}]^i)$$

$$\delta([s_{k_{i-1}+1} \ldots s_{k_i}]^i) =$$

$$(Length([s_{k_{i-1}+1} \ldots s_{k_i}]^i))^w$$

$$\times An * \left(1 - \frac{k_i - k_{i-1} + 1}{Length([s_{k_{i-1}+1} \ldots s_{k_i}]^i)}\right)$$

$$\times \log\left(\sqrt{Fre([s_{k_{i-1}+1} \ldots s_{k_i}]^i)} + 1\right)$$

Word Alignment Based EBMT

- Translation Selection
  - Evaluate the quality of translation segment with word translation probability;
  - And with the number of aligned words in the segment

$$T = \arg\max_{T'} P(T' \mid S) * P(An \mid m, l)$$

Contents

- Background
- Word Alignment Based EBMT
- Experiments and Discussions
  - Data settings
  - RBMT—the rival system
  - Performance and discussions
- Conclusion

Experiments and Discussions

- Data settings
  - Supplied: 20,000 beads for training;
  - Un-restricted: extra 58,600 beads including dining, traffic, sports and travelling domain;
  - Chinese-English dictionary:88,378 entries, for Chinese word segmentation and default translation;
  - Tested on the development corpus and the final test set;

Experiments and Discussions

- RBMT—the rival system
  - A typical Chinese-English translation system based on “analysis-transfer-generation”;
  - First implemented as “BT863” in 1995, top system in MT evaluation held by National 863 project;
  - Re-implemented in 1999-2000, with solid improvement in Chinese analysis;
  - Integrated with Head-driven Chinese parser in 2002;
  - Rule base optimization in 2003;

Experiments and Discussions

- Performance: development corpus

		BLEU-4	NIST-5
EBMT	Supplied -Optimal	0.2082	5.5754
	Supplied - Baseline	0.2052	5.3975
	Un-restricted -Optimal	0.2209	5.5940
	Un-restricted - Baseline	0.2236	5.6220
RBMT		0.1477	5.1990

Experiments and Discussions

- Performance: final result

	Supplied		Un-restricted	
	Optimal	Baseline	Optimal	Baseline
BLEU4	0.2099	0.2113	0.2438	0.2427
NIST5	5.9554	5.927	6.1354	6.0603
GTM	0.6013	0.5988	0.6119	0.6152
WER	0.6169	0.6112	0.5941	0.5906
PER	0.5003	0.4976	0.4872	0.4820

Experiments and Discussions

- Discussions
  - Performance of word alignment tool:
    - 80% on F-measure for both general and computer domain bilingual corpus[Yajuan et al, 2001]
  - Extended parallel examples are, linguistically, noise;
  - Locked example sometimes is a whole sentence.
  - No essential generation processing like: reordering and inflection

Conclusion

- A bi-direction CE EBMT:
  - Requires only a word aligned Chinese English bilingual corpus;
  - Example extraction efforts purely based on word alignment;
  - Our approach optperforms a well built RBMT system;
- A prototype, promising but need detailed polish!

Thanks !