New Word & Phrasal Alignment Methods for Machine Translation

Yanjun Ma, Nicolas Stroppa, Andy Way

NCLT, School of Computing,
Dublin City University,
Dublin 9, Ireland

Outline

- Bootstrapping Word Alignment via Word Packing (ACL 07)
- Alignment-Guided Chunking (TMI 07)
- Hybrid Chunking for Alignment (Master Thesis 06, Tsinghua Univ.)

Motivation

- What ?
  - One ‘Word’ Aligned to a Sequence of ‘Words’

抱歉: excuse me	fifteen: 十 五
报警: call the police	flight: 次 航班
杯: cup of	get: 拿 到
必须: have to	here: 在 这里

- Why ?
  - Word Alignment for SMT
  - Reduce Word Alignment Complexity
  - Bilingual Tokenization to Bridge Translational Divergences?

Word Packing

- Candidate Extraction
  - Perform 1:*n* Word Alignment
- Candidate Reliability Estimation
  - Re-estimate the Word Alignment
- Bootstrapping Word Alignment via Word Packing
  - Searching for the Best Parameters (Word Packing Scheme) Leading to Best MT Output

Word Packing (Cont'd)

- Candidate Extraction
  - Focus on *1:n* (*n* >1)alignment

    $a_i =< c_i, E_i >$
  - Alignment Models
    - IBM Fertility Models
    - HMM Word and Phrase Alignment Model (Deng & Byrne 2005)
    - N-gram Model (Lv 2003) etc.
  - Packing Configuration
    - Consecutive Sequence of Words
    - Jump Model ? Packing Inconsecutive Words

Word Packing (Cont'd)

- Candidate Reliability Estimation
  - Co-occurrence Frequency:

$$COOC(c_i, E_i)$$

  - Alignment Confidence

$$AC(a_i) = \frac{C(a_i)}{COOC(c_i, E_i)}$$

  - Dice Coefficient

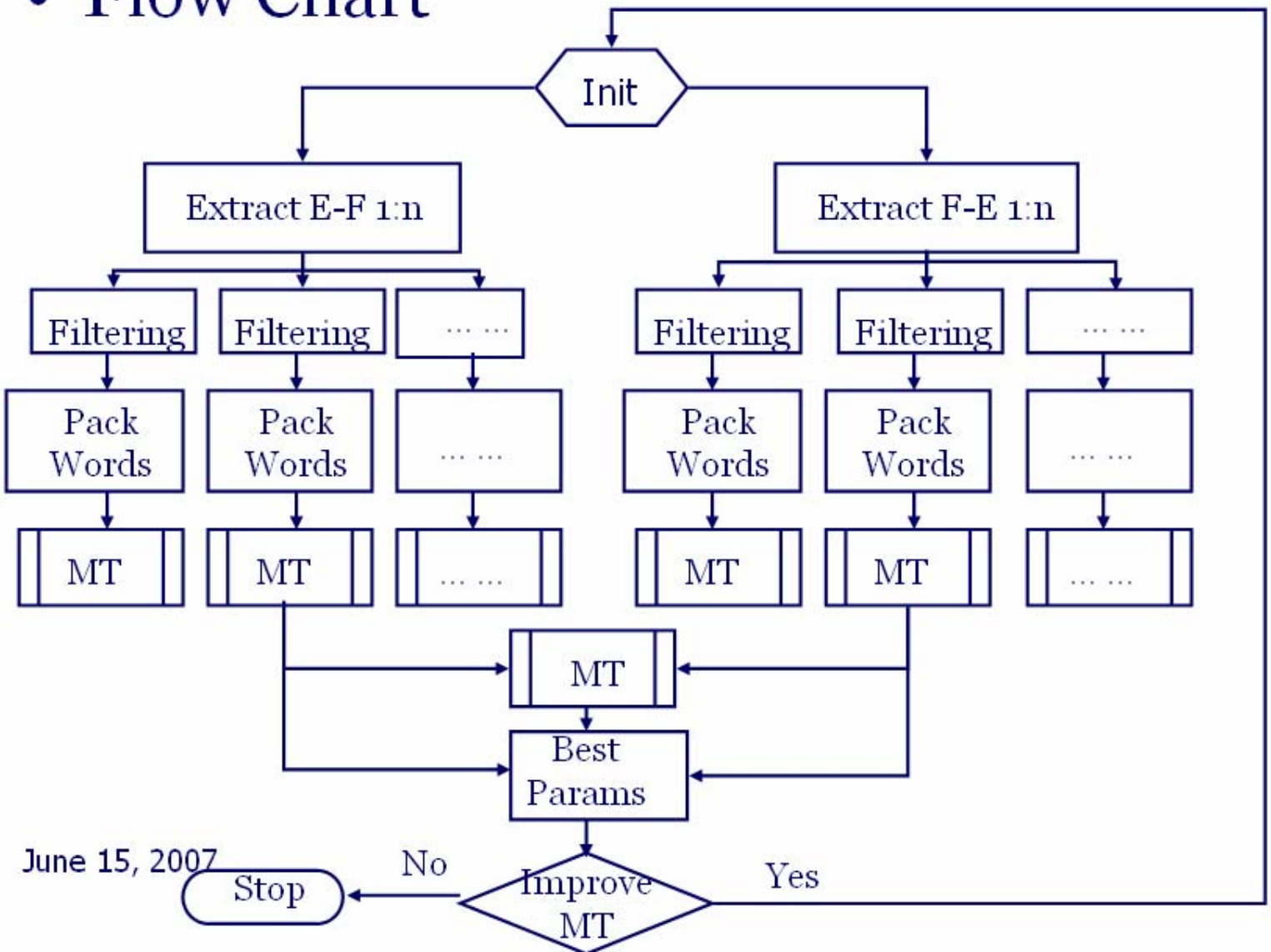
$$Dice = \frac{2 * p(a_i)}{p(c_i) + p(E_i)}$$

Word Packing (Cont'd)

- Bootstrapping Word Alignment
  - Parameters
    - 1:*n*, Maximum of *n*
    - Reliability Estimation
      - COOC, AC, Dice, Chi-Square, etc.
    - Maximum Packing Step
  - Search for the Best Parameters on Development Data

Word Packing (Cont'd)

- Flow Chart



Experiments & Analysis

- Evaluation: MT Quality
- Experimental Results on IWSLT 2006 Chinese-English Task

	BLEU	WER	PER
Baseline	18.55	71.39	53.48
WP. k=1	19.02	69.81	52.14
WP. k=2	19.45	69.59	52.41

Experiments & Analysis (Cont'd)

- What has been changed after word packing ? Where does the improvement come from ?
  - Vocabulary & Length
  - Phrase Table Reconstruction
  - Investigating Histograms of Decoding ?

Experiments & Analysis (Cont'd)

- Exploiting Different Word Segmentations

	BLEU	WER	PER
Baseline (Manual Seg.)	18.55	71.39	58.48
LDC Seg.	16.76	76.17	57.59
HIT Seg.	16.84	71.05	54.59
Baseline + WP	**19.45**	**69.59**	**52.41**
LDC + WP	17.33	73.83	56.25
HIT + WP	17.58	71.69	54.88

Experiments & Analysis (Cont'd)

- Do we need Chinese Word Segmentation for Statistical Machine Translation? (Xu et al. 2006)

	BLEU	WER	PER
Baseline (Manual Seg.)	18.55	71.39	58.48
No Segmentation	18.51	70.83	51.83

Experiments & Analysis (Cont'd)

- Does it Work for other Language Pairs?

Language Pairs	Data	Work ?
Chinese-English	IWSLT 2006	Yes
Japanese-English	IWSLT 2006	Yes ?
Arabic-English	IWSLT 2006	No

Discussion

- Hard Balance between Data Sparseness and *1:n* Alignment Complexity Reduction
  - Pack too many Words: Noise & Data Sparseness (cf. Arabic, Czech, German etc.)
- Extraction and Re-estimation of Packed Words: Context Sensitive
  - Risk of Re-tokenization
    - Make Decision based on Context

Discussion (Cont'd)

- ONLY Packing and NO Unpacking
  - Doom of Wrongly Packed Words
- Time-Consuming

Outline

- Bootstrapping Word Alignment via Word Packing (ACL 07)
- Alignment-Guided Chunking (TMI 07)
- Hybrid Chunking for Alignment (Master Thesis 06, Tsinghua Univ.)

Motivation

- What ?
  - Monolingual Chunking in Bilingual Context
- Why ?
  - Monolingual Chunking in Monolingual Context
    - Based on Hand-crafted Grammar
    - No Bilingual Awareness

Motivation

- Alignment-Guided Chunking
  - Monolingual Chunking in Bilingual Context

Alignment-Guided Chunking

- Example
  - Word Alignment

Cette ville est chargée de symboles puissants pour les trois religions monothéistes .

The city bears the weight of powerful symbols for all three monotheistic religions .

  - AGC Chunks

Cette ||| ville ||| est ||| chargée ||| de ||| symboles puissants ||| pour ||| les ||| trois ||| religions monothéistes ||| .

The ||| city ||| bears ||| the weight ||| of ||| powerful symbols ||| for ||| all ||| three ||| monotheistic religions ||| .

Multi-level Chunking

- Chunking: Ranking V.S. Classification

The ||| city ||| bears ||| the ||| weight ||| of ||| powerful ||| symbols
0.7069 0.5307 0.5467 0.4527 0.3777 0.4098 0.4162
||| for ||| all ||| three ||| monotheistic ||| religions ||| .
0.4318 0.4253 0.3807 0.5655 0.5078 0.9796

- Probability Estimation
  - Various Machine Learning Techniques

Potential Usage of AGC in SMT

- Decoding
  - Provide Prior Knowledge on Sentence Chunking
- AGC Chunks as a Parameter in Decoding ?

Outline

- Bootstrapping Word Alignment via Word Packing (ACL 07)
- Alignment-Guided Chunking (TMI 07)
- Hybrid Chunking and Chunk Alignment (Master Thesis 06, Tsinghua Univ.)

Motivation

- What ?
  - Chunking to Facilitate Chunk Alignment
- Why ?
  - Handle NULL Word Alignments (Function Words)
  - Reduce Complexity of Word Alignment via Chunk Boundaries (Sun et al. 2000)
  - Incorporate Syntax Information: Marker Words and Base Noun Phrases

Hybrid Chunking

- Finding the Best Chunking Approach for Bilingual Corpora
  - Criteria
    - Based on State-of-the-art Chunking Strategies
    - Facilitate Word Alignment (avoiding *n*:*n* chunk alignment)
    - Incorporate Syntax Information
  - Case Study: Chinese-English
    - Combining Marker-based Chunking and Base Noun Phrase Identification
      - Pre-defined Sets of Marker Words
      - Various Machine Learning Techniques

Chunk Alignment

- Anchor Word Alignment
  - Reliable Word Alignment Information
  - Anchor Chunk Alignment
- Distance Distortion for Disambiguation
  - Based on Anchor Chunk Alignment

Discussion

- Integrate Chunking and Chunk Alignment ?
- Syntax-rule Derivation based on Chunked Bilingual Corpus ?
- Testing in MT Systems

Outline

- Bootstrapping Word Alignment via Word Packing (ACL 07)
- Alignment-Guided Chunking (TMI 07)
- Hybrid Chunking for Alignment (Master Thesis 06, Tsinghua Univ.)

Conclusion ?

- Investigating Word and Phrasal Alignment in a Bilingual Context
  - What is a Word ?
  - What is a Chunk ?

Thanks!

Comments & Questions?