Word Segmentation and their Integration in Machine Translation

Advanced MT Seminar

ThuyLinh Nguyen

thuylinh@cs.cmu.edu

Word Segmentation Problems

日文章魚怎麼說？

'How do you say octopus in Japanese?'

Plausible Segmentation

日文	章魚	怎麼	說
ri4-wen2	*zhang1-yu2*	*zen3-me0*	*shuo1*
'Japanese'	'octopus'	'how'	'say'

Implausible Segmentation

日	文章	魚	怎麼	說
ri4	*wen2-zhang1*	*yu2*	*zen3-me0*	*shuo1*
'Japan'	'essay'	'fish'	'how'	'say'

Word Segmentation for MT

- Use word segmentation toolkit to segment character sequences into words before the training and translation.
- Each Chinese character is interpreted as a single word and learn the segmentation from Chinese character - English word alignment. (Xu et al. [2004])
- Confusion networks: Take different segmentations into account and represent them as lattice. The input of the translation system is a set of lattices. (Xu [2005])

Word Segmentation Problems

- Ambiguity
  - A character can be a word component in one context or a word by itself in other context.
  - A character can occur in different positions.

Position	Example
Left	产生 'to come up with'
Word by itself	产小麦 'to grow wheat'
Middle	生产线 'assembly line'
Right	生产 'to produce'

Word Segmentation Problems

- Ambiguity
  - A character can be a word component in one context or a word by itself in other context.
  - A character can occur in different positions.
- Unknown words
  - New words are combinations of existing words.
  - Names are created by combining characters in unpredictable manner.
  - Transliteration of foreign names.

Word Segmentation Problems

- Ambiguity
  - A character can be a word component in one context or a word by itself in other context.
  - A character can occur in different positions.
- Unknown words
  - New words are combinations of existing words.
  - Names are created by combining characters in unpredictable manner.
  - Transliteration of foreign names.
- There is no widely accepted definition of Chinese word. (Sproat et al. [1994])used 6 people segmented the same text. The segmentation consistency is only 76%.

Word Segmentation methods

- Purely dictionary-based approach (Cheng et al. [1999])
  - Address the ambiguity problem with maximum matching heuristic.
  - Pros: Simple, good heuristic.
  - Cons: Depends on the coverage of the dictionary.

Word Segmentation methods

- Purely dictionary-based approach (Cheng et al. [1999])
  - Address the ambiguity problem with maximum matching heuristic.
  - Pros: Simple, good heuristic.
  - Cons: Depends on the coverage of the dictionary.
- Purely statistical-based approach
  - Use Point-wise mutual information or EM.
  - Pros: Not depend on a dictionary.
  - Cons: Low accuracy.

Word Segmentation methods

- Purely dictionary-based approach (Cheng et al. [1999])
  - Address the ambiguity problem with maximum matching heuristic.
  - Pros: Simple, good heuristic.
  - Cons: Depends on the coverage of the dictionary.
- Purely statistical-based approach
  - Use Point-wise mutual information or EM.
  - Pros: Not depend on a dictionary.
  - Cons: Low accuracy.
- Statistical-based approach using manual word segmentation data.

CRF for Word Segmentation

Peng et al. [2004] & Tseng et al. [2005]

- Word segmentation as Character Tagging problem

日	文	章	魚	怎	麼	說
S	NS	S	NS	S	NS	S NS

CRF for Word Segmentation

Peng et al. [2004] & Tseng et al. [2005]

- Word segmentation as Character Tagging problem

日	文	章	魚	怎	麼	說	
S	NS	S	NS	S	NS	S	NS

- Conditional Random Field model
  Let $\mathbf{c} = (c_1, c_2, \ldots, c_K)$ be a Chinese sentence,
  $\mathbf{t} = (t_1, t_2, \ldots, t_K)$ be the character tags of $\mathbf{c}$.

$$\Pr(\mathbf{t}|\mathbf{c}) = \frac{1}{Z(\mathbf{c})} \exp\left(\sum_{k=1}^{k=K} \sum_i \lambda_i f_i(t_{k-1}, t_k, \mathbf{c}, k) \right)$$

CRF for Word Segmentation

- Unknown words detection
  - Peng et al. [2004]: Use forward backward algorithm to calculate the confidence of word segment.
  - Tseng et al. [2005]: Add additional features to the model i.e the first and the last characters of rare words.

CRF for Word Segmentation

- Unknown words detection
  - Peng et al. [2004]: Use forward backward algorithm to calculate the confidence of word segment.
  - Tseng et al. [2005]: Add additional features to the model i.e the first and the last characters of rare words.
- Results

Sighan Bakeoff 2003	*F-score Tseng et al (2005)*	*F-score Peng et al. (2004)*
CTB	0.863	0.849
AS	0.970	0.956
HK	0.947	0.928
PK	0.953	0.941

Do We Need Word Segmentation for SMT

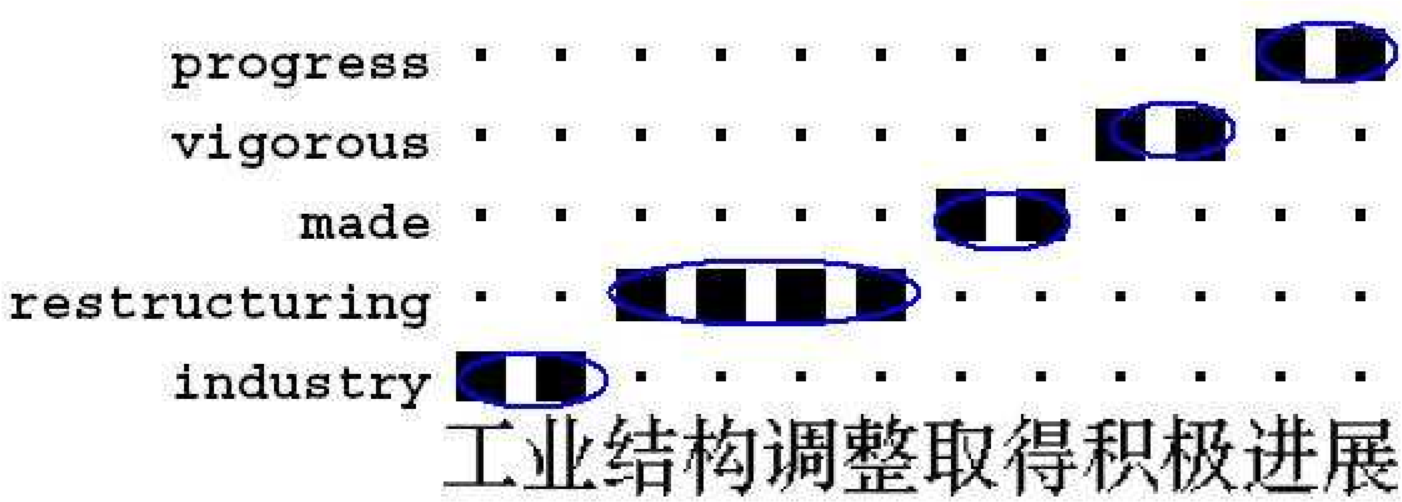
Xu et al. [2004]

- Each Chinese character is interpreted as one “word”.
- Aligned Chinese characters with English text.

Do We Need Word Segmentation for SMT

Xu et al. [2004]

- Each Chinese character is interpreted as one “word”.
- Aligned Chinese characters with English text.



- Generate a Chinese word dictionary.
- Use self-learned dictionary for Chinese word segmentation.

Do We Need Word Segmentation for SMT

Word length statistics

word length	LDC dictionary		learned dictionary	
	frequency	[%]	frequency	[%]
1	2 334	18.6	2 368	16.9
2	8 149	65.1	5 486	39.2
3	1 188	9.5	1 899	13.6
4	759	6.1	2 084	14.9
5	70	0.6	791	5.7
6	20	0.2	617	4.4
7	6	0.0	327	2.3
$\geq$8	11	0.0	424	3.0
total	12 527	100	13 996	100

Do We Need Word Segmentation for SMT?

		Chinese	English
Train	Sentences	4 172	
	Characters	172 874	832 760
	Words	116 090	145 422
	Char. Vocab.	3 419 + 20	26 + 20
	Word Vocab.	9 391	9 505
Test	Sentences	993	
	Characters	42 100	167 101
	Words	28 247	26 225

method	error rates		accuracy
	WER	PER	BLEU
no segment.	73.3	56.5	27.6
learned segment.	70.4	54.6	29.1
LDC segment.	71.9	54.4	29.2

Integrated Word Segmentation in SMT

Xu [2005]

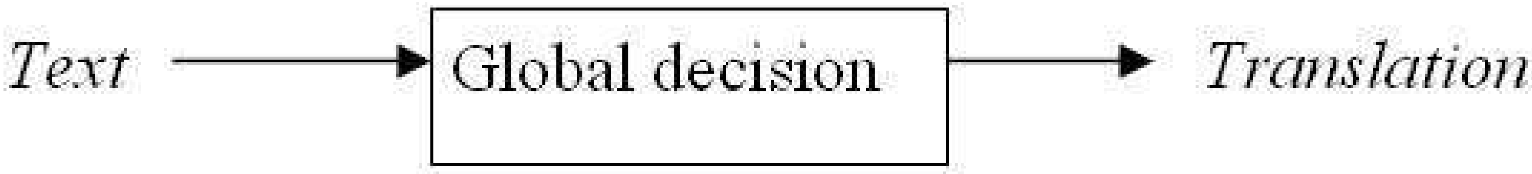
Single best segmentation translation

$$
\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{1}}^{\hat{\mathbf{J}}} &= \arg\max_{\mathbf{f}_{\mathbf{1}}^{\mathbf{J}},\mathbf{J}} \left\{ \Pr\left(\mathbf{f}_{\mathbf{1}}^{\mathbf{J}} | \mathbf{c}_{\mathbf{1}}^{\mathbf{K}}\right) \right\} \\
\hat{\mathbf{e}}_{\mathbf{1}}^{\hat{\mathbf{I}}} &= \arg\max_{\mathbf{e}_{\mathbf{1}}^{\mathbf{I}},\mathbf{I}} \left\{ \Pr\left(\mathbf{e}_{\mathbf{1}}^{\mathbf{I}} | \hat{\mathbf{f}}_{\mathbf{1}}^{\hat{\mathbf{J}}}\right) \right\}
\end{aligned}
$$

Integrated Word Segmentation in SMT

Xu [2005]

Segmentation lattice translation

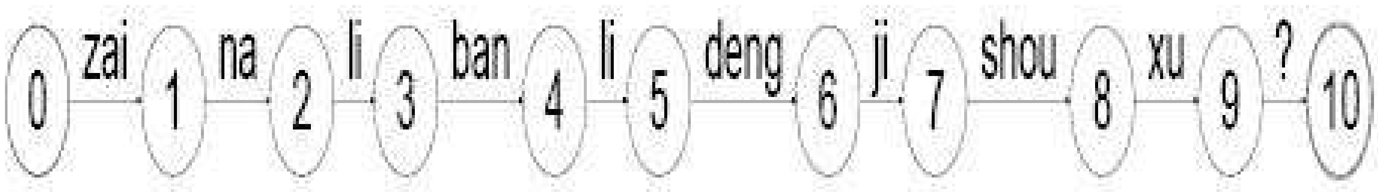


$$\begin{aligned}
\hat{e}_1^{\hat{I}} &= \operatorname*{argmax}_{I,e_1^I} \left\{ Pr(e_1^I | c_1^K) \right\} \\
&= \operatorname*{argmax}_{I,e_1^I} \left\{ \sum_{f_1^J} Pr(f_1^J, e_1^I | c_1^K) \right\} \\
&= \operatorname*{argmax}_{I,e_1^I} \left\{ \sum_{f_1^J} Pr(f_1^J | c_1^K) \cdot Pr(e_1^I | f_1^J, c_1^K) \right\} \\
&\cong \operatorname*{argmax}_{I,e_1^I} \left\{ \max_{f_1^J} \left\{ Pr(f_1^J | c_1^K) \cdot Pr(e_1^I | f_1^J) \right\} \right\}
\end{aligned}$$

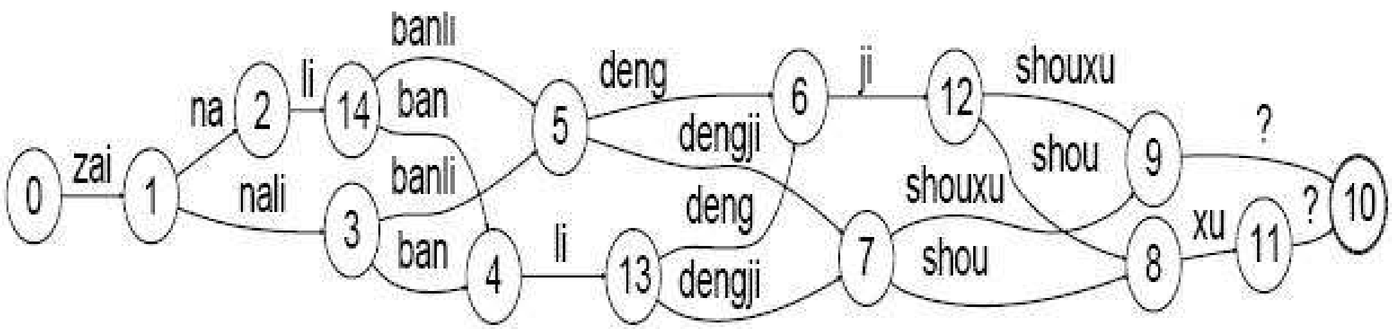
Integrated Word Segmentation in SMT

Xu [2005]

- Input sentence at the character level



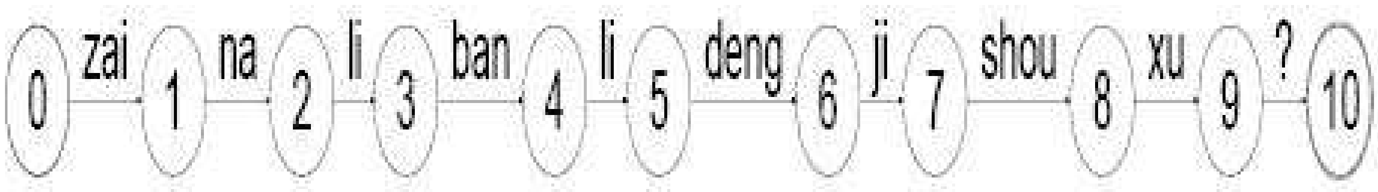
- Segmentation lattice



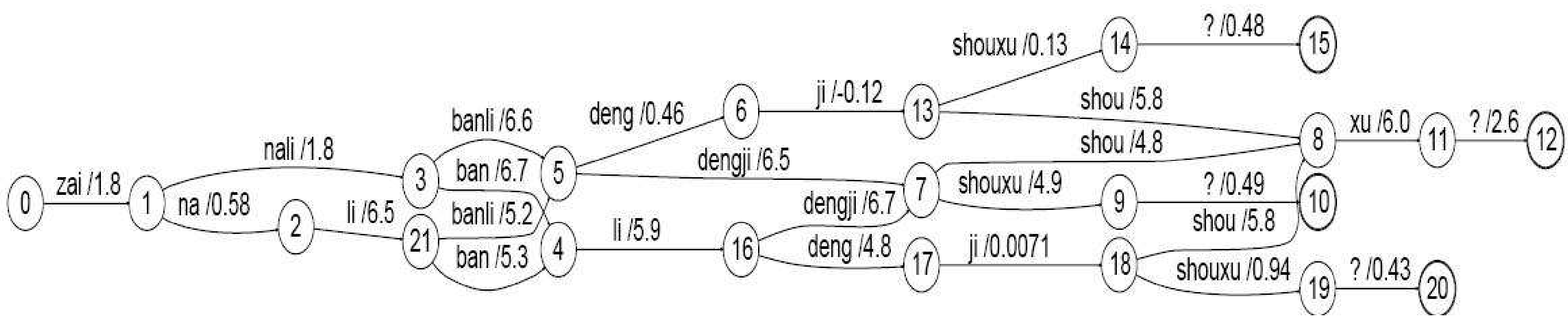
Integrated Word Segmentation in SMT

Xu [2005]

- Input sentence at the character level



- Segmentation lattice with weights



Integrated Word Segmentation in SMT

Xu [2005]

Corpus statistics

		Chinese		English
Train:	Sentences	19 851		
	Running Words	18 1247		159 655
	Vocabulary	7 610		6 955
	Singletons	3 512		2 938
CStar'03:	Sentences	506		
		Words	Characters	Words
	Running Words/Characters	3 515	4 757	65 604
	Vocabulary	870	800	2 078
	OOVs (running words/characters) [%]	5.40	8.74	14.3
	OOVs (in vocabulary) [%]	18.4	26.3	20.6

Integrated Word Segmentation in SMT

Translation results

- Monotone finite state transducer

Segmentation methods	WER [%]	PER [%]	NIST	BLEU [%]
Single-best (manual) segmentation	51.3	43.1	3.60	28.5
Segmentation lattice without weights	51.6	42.2	4.69	29.0

- Phrase based system

Segmentation methods	WER [%]	PER[%]	NIST	BLEU[%]
Single-best (manual) segmentation	53.6	43.8	8.18	38.9
Segmentation lattice without weight	47.0	38.1	8.09	40.2
Segmentation lattice with bi-gram LM	47.2	38.0	8.18	40.4

Conclusion & Discussion

- Very few research on word segmentation for machine translation

Conclusion & Discussion

- Very few research on word segmentation for machine translation
- GIZA++ can produce error alignments.

Conclusion & Discussion

- Very few research on word segmentation for machine translation
- GIZA++ can produce error alignments.
- Unalignment of English words and Chinese characters.

Conclusion & Discussion

- Very few research on word segmentation for machine translation
- GIZA++ can produce error alignments.
- Unalignment of English words and Chinese characters.
- Word reordering problems.

References

K. S. Cheng, G. H. Young, and Wong. A study on word-based and integral-bit chinese text compression algorithms. *Journal of the American Society for Information Science*, 50(3):218–228, 1999.

Fuchun Peng, Fangfang Feng, and Andrew Mccallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of Coling 2004*, pages 562–568, Geneva, Switzerland, Aug FebruaryMarch–Aug FebruaryJuly 2004. COLING.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. A stochastic finite-state word-segmentation algorithm for chinese. In *Meeting of the Association for Computational Linguistics*, pages 66–73, 1994. URL #.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. A conditional random field word segmenter. 2005. URL `http://www.aclweb.org/anthology-new/W/W06/`

Xu. Integrated chinese word segmentation in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 141–147, Pittsburgh, PA, October 2005.

J. Xu, R. Zens, and H. Ney. Do we need chinese word segmentation for statistical machine translation? In *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, pages 122–128, Barcelona, Spain, July 2004.