

Integrated Chinese Word Segmentation in Statistical Machine Translation

Jia Xu, Evgeny Matusov, Richard Zens, Hermann Ney

**Human Language Technology and Pattern Recognition
Chair of Computer Science VI
RWTH Aachen University
Germany**

Content

- 1. problem description**
- 2. introduction**
- 3. baseline statistical machine translation system**
- 4. word segmentation model and segmentation approaches**
- 5. computational steps**
- 6. corpus statistics**
- 7. translation results and computational requirements**
- 8. conclusion**

Problem Description

Words are not separated by white space in Chinese sentence

Standard approach in statistical machine translation:

- **segmentation of Chinese character sequences into words**
- **training and translation are performed afterwards**

The problems of the standard approach:

- 1. segmentation may contain errors**
- 2. for a given character sequence,
the best segmentation depends on its context**
- 3. manual segmentation is not necessarily the best segmentation for translation
(e. g. into English)**

Introduction

Translation at word/character level [Xu et al., 2004]

1. word level: training and test texts are segmented using a segmentation tool → usually better results
2. character level: the texts are used unsegmented

Integrated segmentation:

- different segmentation alternatives instead of a single segmentation are taken into account
- the segmentation process is integrated with the search for the best translation

Advantages of the integrated segmentation:

- improvement in translation quality
- the Chinese text can be translated at character level

Baseline SMT System

Given:

- a source sentence $f_1^J = f_1 \dots f_j \dots f_J$ to be translated into
- a target sentence $e_1^I = e_1 \dots e_i \dots e_I$

Choose the target sentence with the highest posterior probability:

$$\begin{aligned}\hat{e}_1^I &= \operatorname{argmax}_{e_1^I} \{Pr(e_1^I | f_1^J)\} \\ &= \operatorname{argmax}_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J | e_1^I)\}\end{aligned}$$

- language model $Pr(e_1^I)$
- translation model $Pr(f_1^J | e_1^I)$

Word Segmentation Model

Given:

- a Chinese source sentence at the character level $c_1^J = c_1 \dots c_k \dots c_K$

The best segmented sentence $\hat{f}_1^{\hat{J}}$ with \hat{J} words can be represented as:

$$\begin{aligned} \hat{f}_1^{\hat{J}} &= \operatorname{argmax}_{f_1^J} \{Pr(f_1^J | c_1^K)\} \\ &= \operatorname{argmax}_{f_1^J} \{Pr(c_1^K | f_1^J) \cdot Pr(f_1^J)\} \end{aligned}$$

Two sub-models are used

Word Segmentation Sub-models

1. correspondence of f_1^J and c_1^K

$$Pr(c_1^K | f_1^J) = \begin{cases} 0 & : C(f_1^J) \neq c_1^K \\ 1 & : C(f_1^J) = c_1^K \end{cases}$$

C is the separation of a word sequence into characters

2. source language model at the word level [Luo et al., 1996]:

$$\begin{aligned} Pr(f_1^J) &= \prod_{j=1}^J Pr(f_j | f_1^{j-1}) \\ &\cong \prod_{j=1}^J p(f_j | f_{j-n}^{j-1}) \end{aligned} \quad (1)$$

Segmentation Approaches - 1

- single-best segmentation (standard approach)

$$\hat{e}_1^I = \operatorname{argmax}_{e_1^I} \left\{ \operatorname{Pr}(e_1^I | \hat{f}_1^J) \right\}$$

- segmentation lattice (integrated segmentation)

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

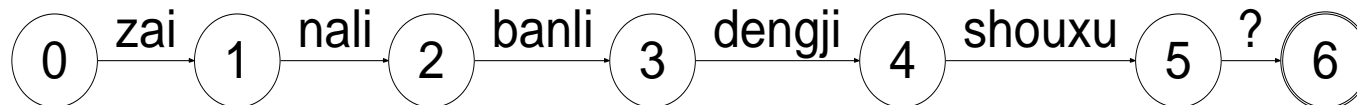
Segmentation Approaches - 2

Example

source sentence in characters:	zai na li ban li deng ji shou xu ?
manually segmented source sentence:	zai nali banli dengji shouxu ?
translation by single-best segmentation:	where to go through boarding formalities ?
translation by segmentation lattice:	where do I make my boarding arrangements ?
reference translation:	where do I complete boarding procedures ?

Single-best segmentation:

- input sentence is a linear automaton



Computational Steps - 1

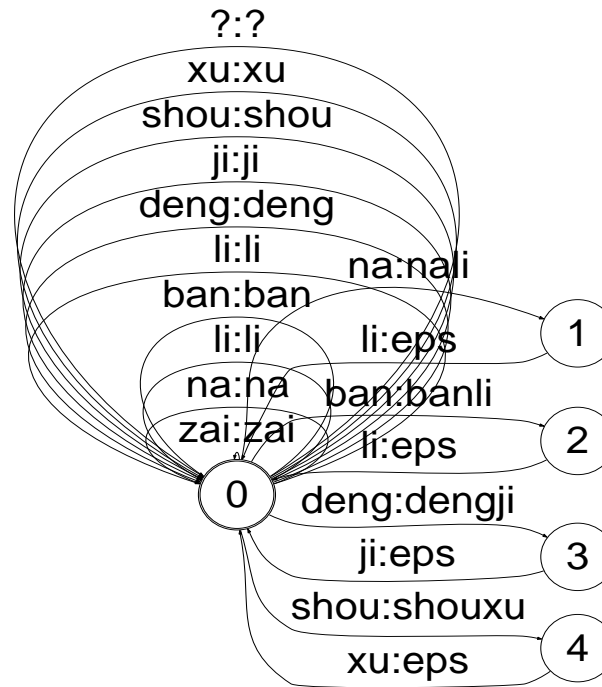
Segmentation lattice

1. generation of the word list from the training corpus vocabulary

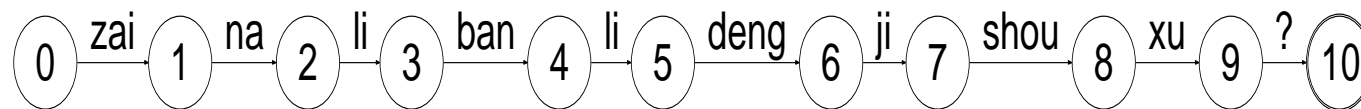
characters	words
zai	zai
..	..
na li	nali
ban li	banli
deng ji	dengji
shou xu	shouxu

Computational Steps - 2

2. conversion of the word list into a segmentation transducer

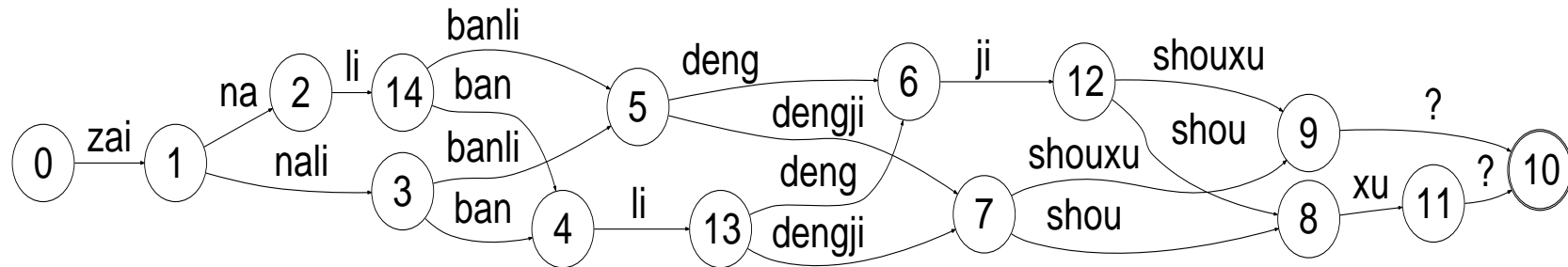


3. the input character sequence is represented as a linear automaton



Computational Steps - 3

4. composition of the linear automaton and the segmentation transducer



5. score the lattice with language model costs

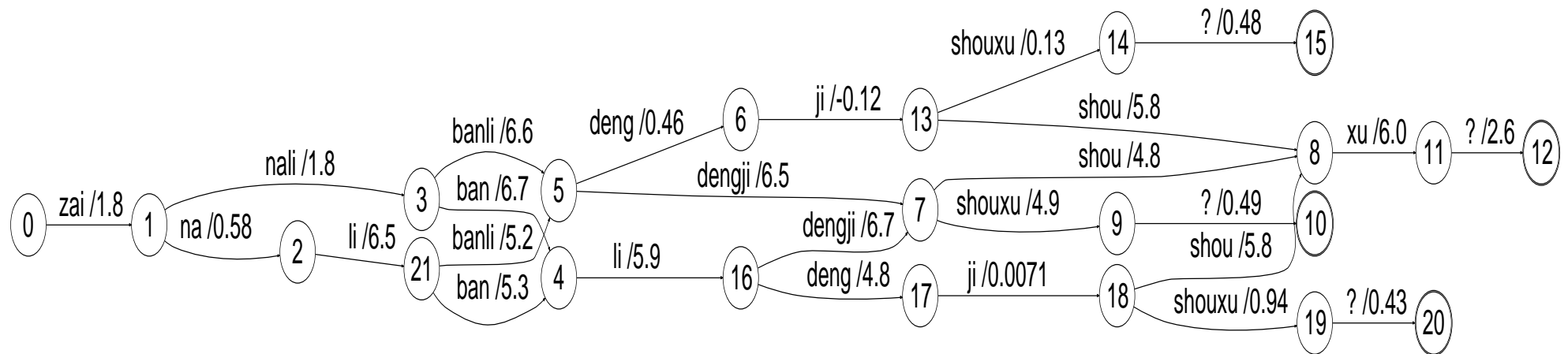
Word segmentation model:

- represents the fluency of a Chinese word sequence
- can be built as an n-gram language model at the word level

Insertion of the word segmentation costs:

- transform the language model into a finite-state transducer
- compose the segmentation lattice with the transducer

Computational Steps - 4



Experimental Results

Integrated segmentation tested on the Basic Travel Expression Corpus (IWSLT 2005 supplied task)

- training and evaluation data with preprocessing
- evaluation set: CStar'03

		Chinese		English
Train:	Sentences	19851		
	Running Words	181247		159655
	Vocabulary	7610		6955
	Singletons	3512		2938
Evaluation:	Sentences	506		
		Words	Characters	Words
	Running Words/Characters	3515	4757	65604
	Vocabulary	870	800	2078
	OOVs (r. words/characters)	190	416	9422

Translation Results

- translation performance with monotone finite-state transducer based translation [Kanthak et al., 2004]

segmentation methods	WER [%]	PER [%]	NIST	BLEU [%]
single-best (manual) segmentation	51.3	43.1	3.60	28.5
unweighted segmentation lattice	51.6	42.2	4.69	29.0

- translation performance with phrase-based translation [Zens et al., 2004]

segmentation methods	WER [%]	PER [%]	NIST	BLEU [%]
single-best (manual) segmentation	53.6	43.8	8.18	38.9
unweighted segmentation lattice	47.0	38.1	8.09	40.2
+ bi-gram LM scores	47.2	38.0	8.18	40.4

Computational Requirements

- **phrase-based translation system**
- **lattice density: the number of arcs in the lattice divided by the number of characters in the sentence**

segmentation approach	Lattice density	Memory [MB]	Speed [sec./sentence]
single-best segmentation	-	54.2	0.26
unweighted segmentation lattice	1.5	56.9	0.27
+ bigram LM scores	3.9	65.8	0.82

Conclusion

The method:

- **integration of the word segmentation decisions directly in translation search**

Advantages:

- **Chinese input text is at character level,
there is no need to segment the text during the preprocessing**
- **using integrated segmentation results in better translation quality
than when using single-best manual segmentation**

Outlook:

- **better Chinese word segmentation in training**

References

- **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, Barcelona, Spain, pp. 122-128, July 2004.**
- **P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, A statistical approach to machine translation, Computational Linguistics, vol. 16, no. 2, pp. 79-85, June 1990.**
- **X. Luo and S. Roukos, An iterative algorithm to build Chinese language models, in Proc. of the 34th annual meeting of the Association for Computational Linguistics, Santa Cruz, California, June 1996, pp. 139-143.**
- **IWSLT, Intl. workshop on spoken language translation home page, 2005, <http://www.is.cs.cmu.edu/iwslt2005/CFP.html>.**
- **S. Kanthak and H. Ney, FSA: An efficient and flexible C++ toolkit for finite state automata using ondemand computation, in Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, July 2004, pp. 510-517.**
- **R. Zens and H. Ney, Improvements in phrase-based statistical machine translation, in Proc. of the Human Language Technology Conference, Boston, MA, May 2004.**