The TRGTK's System Description for PatentMT at NTCIR-10

Hao Xiong
Torangetek Inc.
Institute of Computing Technology
Chinese Academy of Sciences
2013-6-21

Outline

- Something about TRGTK
- Techniques used in Patent MT
- Summary

Outline

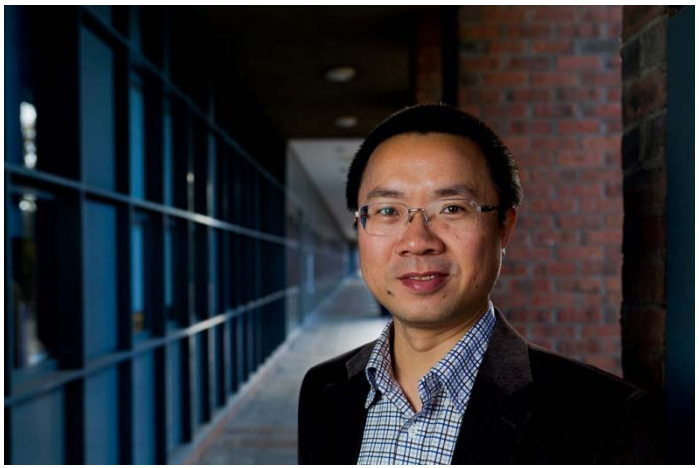
- **Something about TRGTK**
- Techniques used in Patent MT
- Summary

Torangetek Inc.

- Torangetek
  - Translation + Orange + Technology

Torangetek Inc.

- Torangetek
  - Translation + Orange + Technology



Customers

- Spoken Translation Service

- Patent Translation Service

- E-business Translation Service

Outline

- Something about TRGTK
- **Techniques used in Patent MT**
- Summary

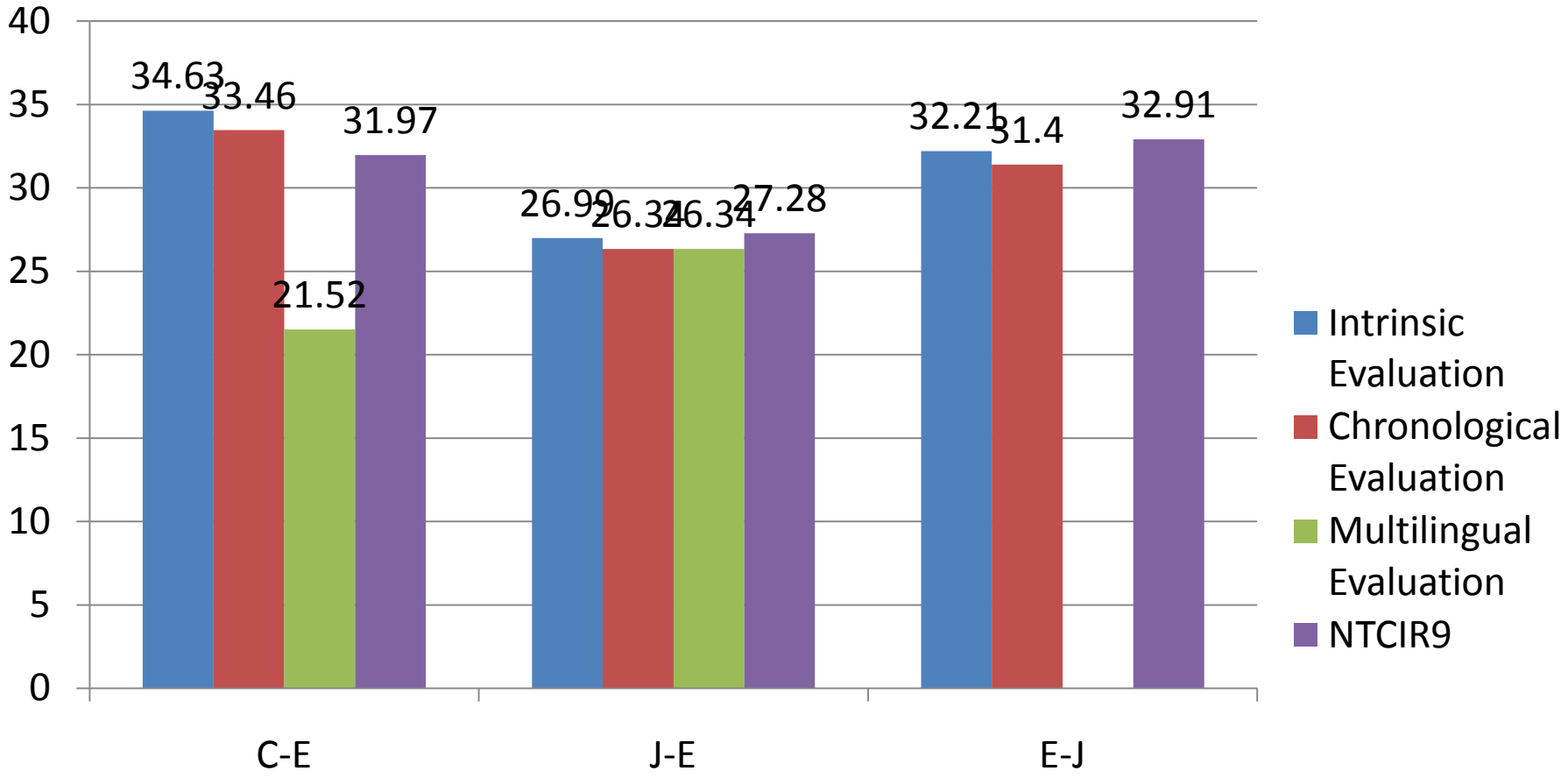
Goal

- **Translation Speed** Versus **Translation Quality**
- NTCIR-9
  - ICT: Academy
  - Quality
- NTCIR-10
  - Torangetek: Company
  - Fast and Stable Translation Service

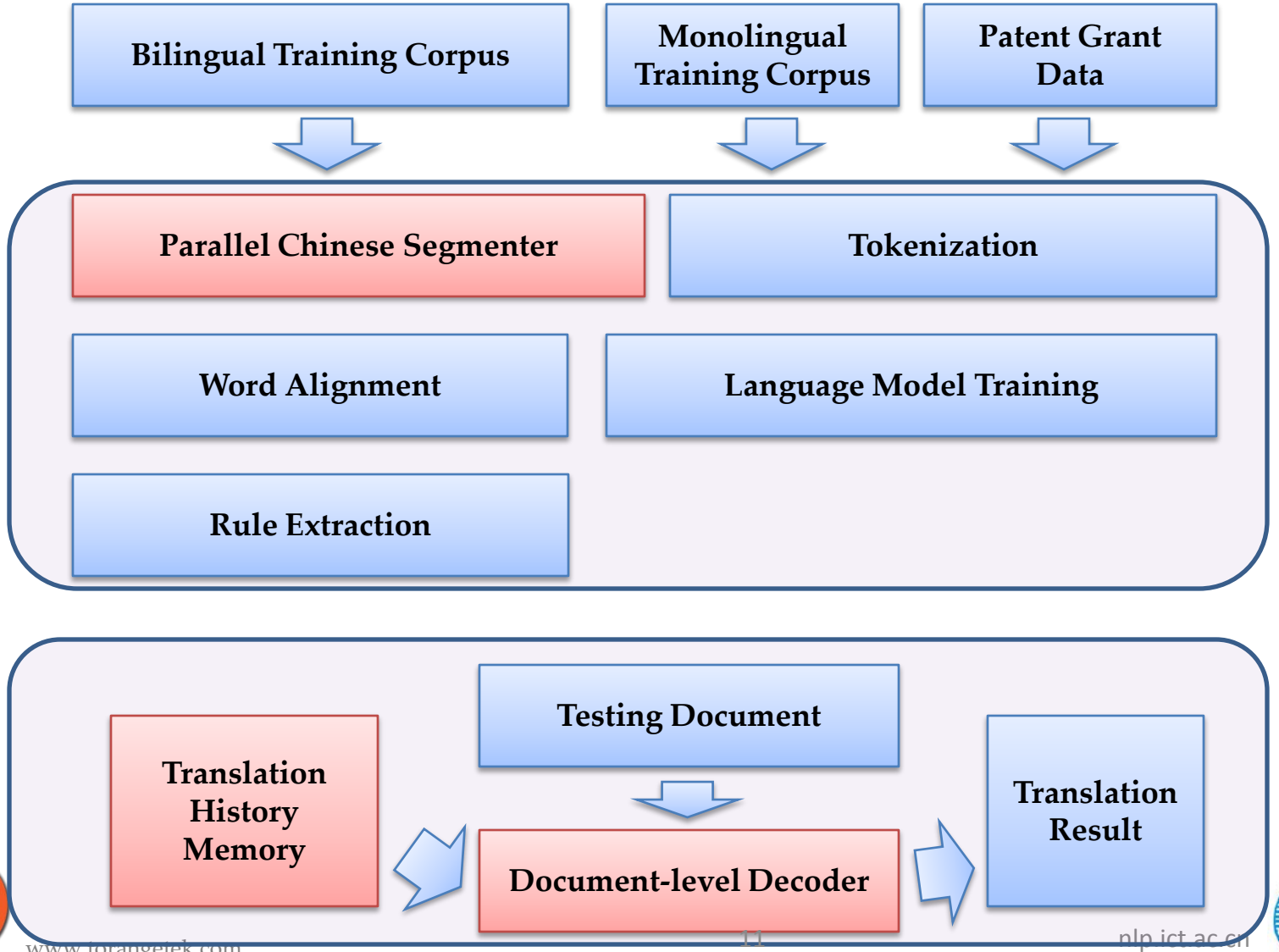
Data Usage

System	Bilingual	Monolingual
C-E	1 Million	40 Million
J-E	3 Million	40 Million
E-J	3 Million	73 Million

Final Results



Overall Architecture



Traditional Chinese Segmenter

- Training is Off-line
  - Time Insensitive
- Testing is On-line
  - $O(n^2)$ for Effective Searching
  - Patent Sentences: more than 10 words
  - Time Sensitive

Parallel Chinese Segmenter

- Character-based Max-Entropy Model
  - Hwee Tou Ng and Jin Kiat Low, 2004
  - Meng et.,al, 2012
  - BMES
  - 13 Features
  - 10k Entities from Baidu-pedia
- Segmentation History Storage
- Searching on GPUs

Segmentation History Storage

- Practical Application
  - Using 3% Words in Training Set
- Character-based Model
  - Context Length:2 characters

Running Example

Training Instance

奥巴马	与	沙龙	举行	了	会谈
Obama	**and**	**Sharon**	**hold**	**a**	**talk**

Running Example

Training Instance

奥巴马	与	沙龙	举行	了	会谈
Obama	**and**	**Sharon**	**hold**	**a**	**talk**

Testing Instance

布	什	与	沙	龙	举	行	了	会	谈
Bush		**and**	**Sharon**		**hold**		**a**	**talk**	

Running Example

Training Instance

奥巴马	与	沙龙	举行	了	会谈
Obama	and	Sharon	hold	a	talk

Testing Instance

布 什	与	沙 龙	举 行	了	会 谈
Bush	and	Sharon	hold	a	talk

String Matching

Running Example

Training Instance

奥巴马 与 沙龙 举行 了 会谈
Obama and Sharon hold a talk

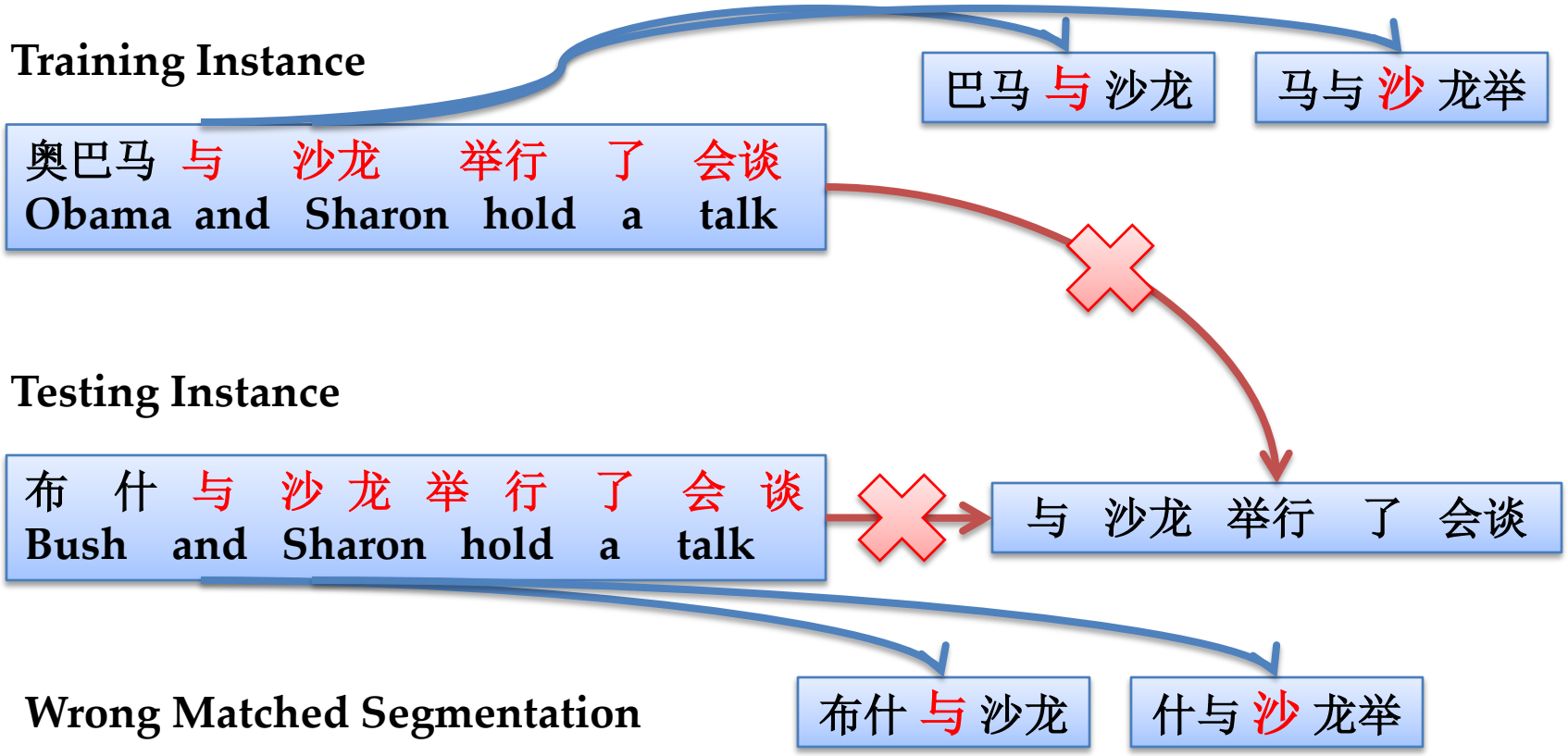
Testing Instance

布 什 与 沙 龙 举 行 了 会 谈
Bush and Sharon hold a talk

与 沙龙 举行 了 会谈

Wrong Matched Segmentation

Running Example



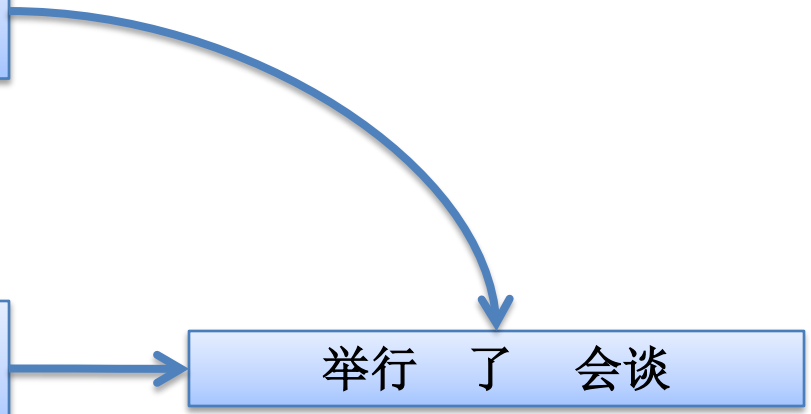
Running Example

Training Instance

奥巴马 与 沙龙 举行 了 会谈
Obama and Sharon hold a talk

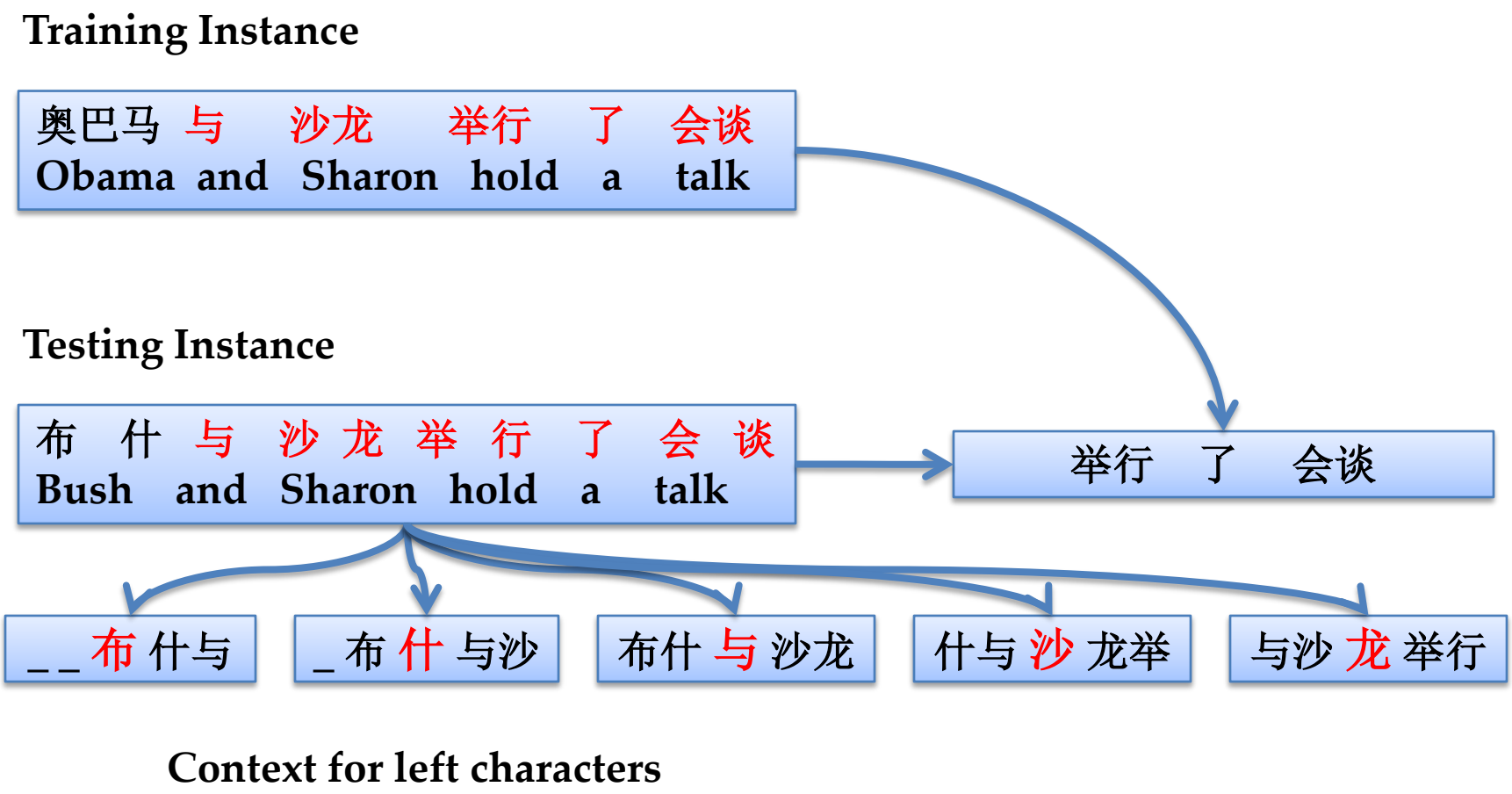
Testing Instance

布 什 与 沙 龙 举 行 了 会 谈
Bush and Sharon hold a talk

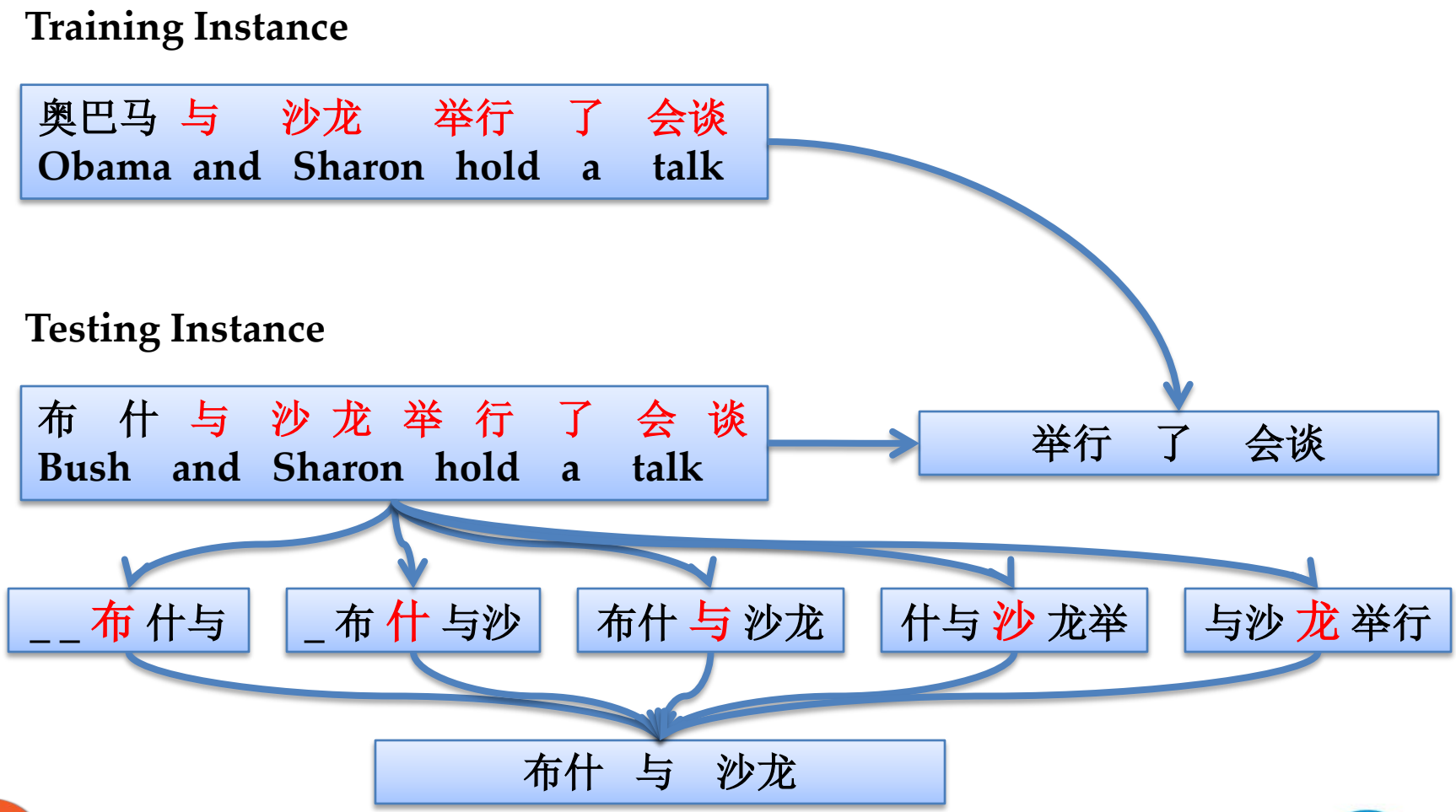


Correct Matched Segmentation

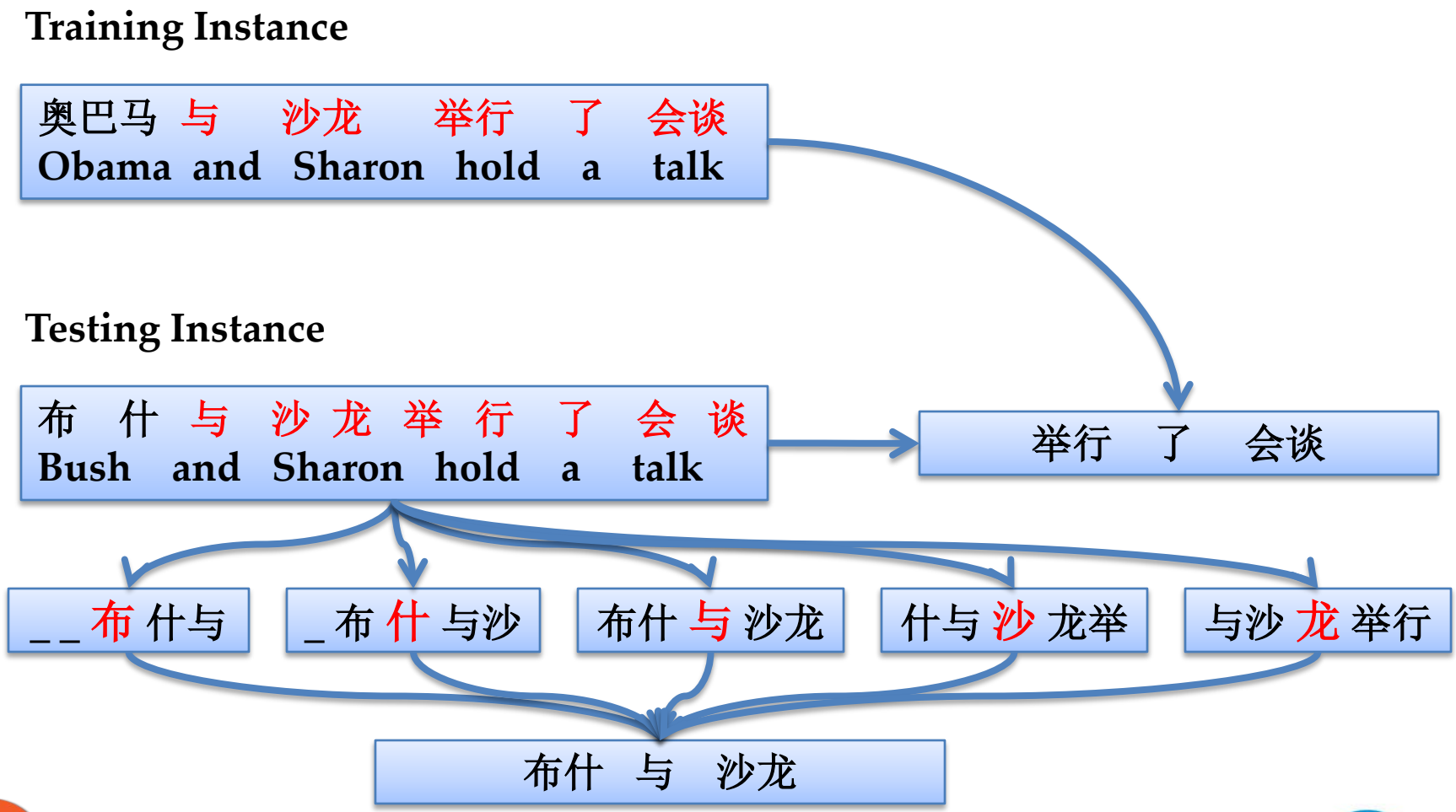
Running Example



Running Example



Running Example



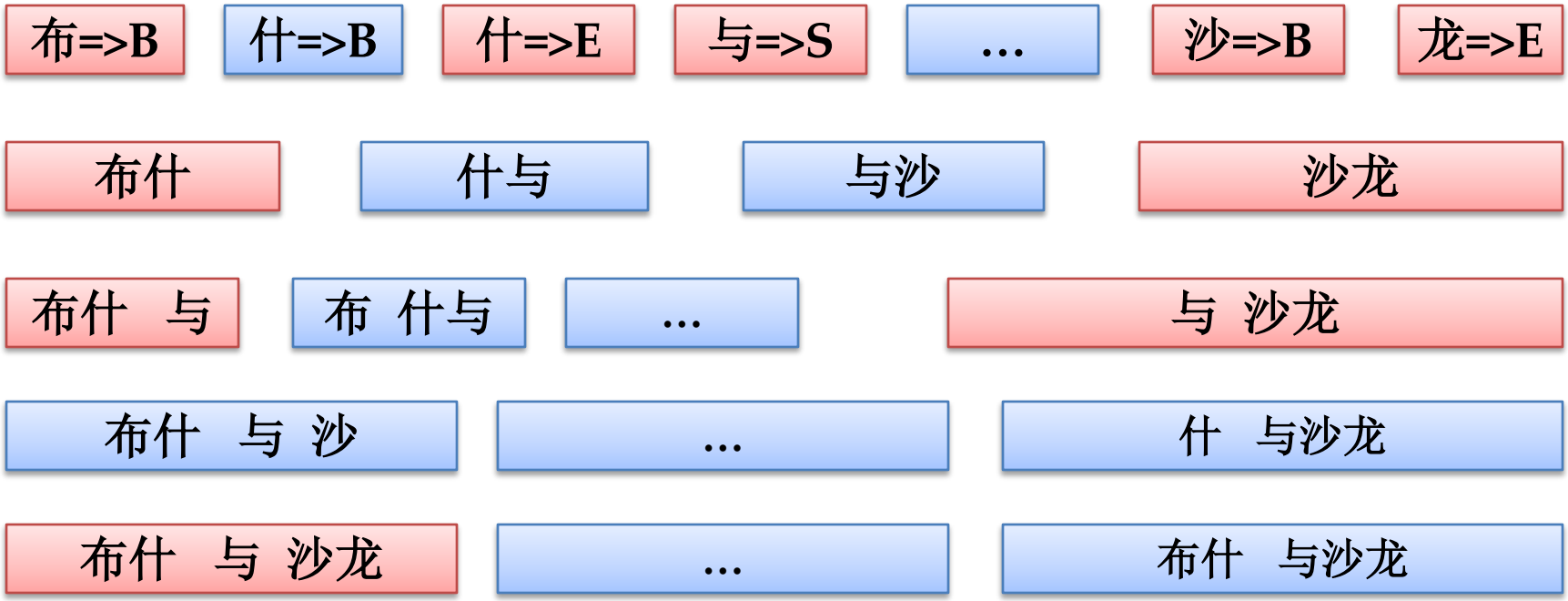
Computing on GPUs

- Inspired by Youngmin Yi et., 2011
  - CKY-Parsing on GPUs
- GPUs
  - NVIDIA GTX480 GPU: 480 processing cores
- Compute Unified Device Architecture

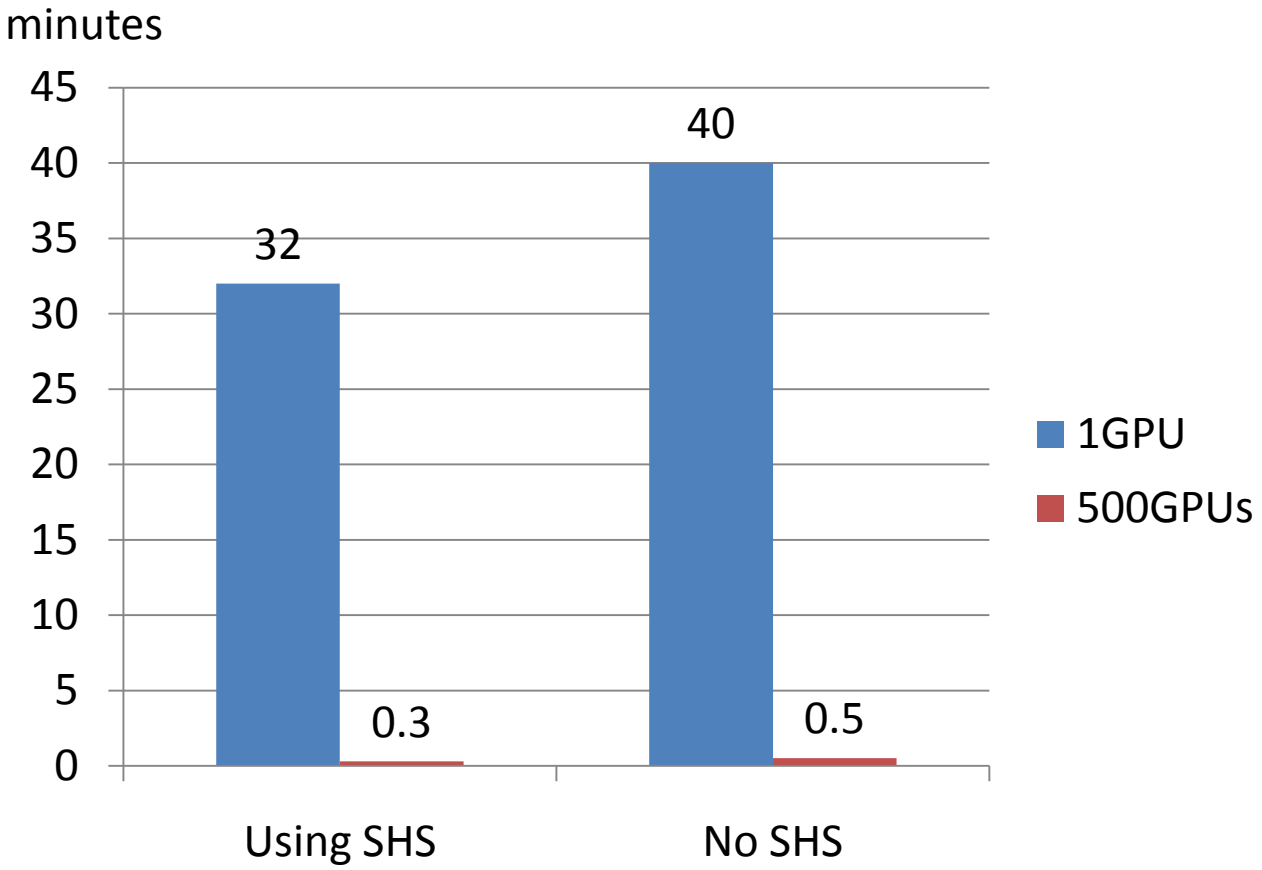
Parallel Computing

- Hierarchical Computing
  - Layer 1: Segmentation with 1 Character
  - Layer 2: Segmentation with 2 Characters
  - …
  - Layer n: Segmentation with n Characters
- Computing in One Layer is Parallel

Parallel Computing



Experiments of Segmentation



SHS: Segmentation Historical Storage(10 Million Extrinsic Patent Chinese Sentences)

Patent Translation

- Document Translation
  - Description of Patents
- Translation Consistency
  - Terminology
- Large Requirements
  - More than 10k docs per day

Document-level Decoder

- Document Generation
  - KNN Algorithm + Distributional Similarity
- Term Recognition
  - C-Value
  - Term Candidates Templates
- **Documental CKY Algorithm**

Documental CKY Algorithm

- Hierarchical Phrase-based Model
- Linking Translation Consistent Cubes
- Shared Score
  - Local Score
    - Translation Prob, Lexical Trans Prob,…, Word Count
  - **Shared Language Model Score**
    - LMs from different contexts in each sentence

Running Example

Sentence 1

$W_1, W_2, .., W_i, .., W_n$

Sentence 2

$W_1, W_2, .., W_j, .., W_n$

Sentence 3

$W_1, W_2, .., W_k, .., W_n$

Running Example

Sentence 1

$W_1, W_2, .., W_i, .., W_n$

Sentence 2

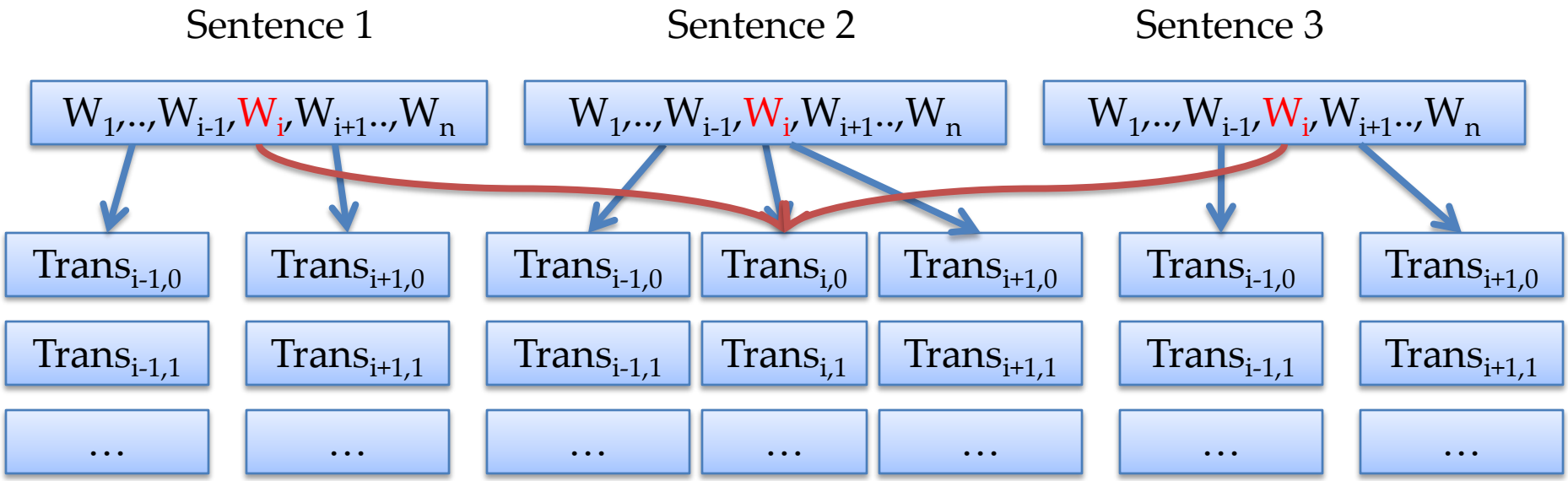
$W_1, W_2, .., W_j, .., W_n$

Sentence 3

$W_1, W_2, .., W_k, .., W_n$

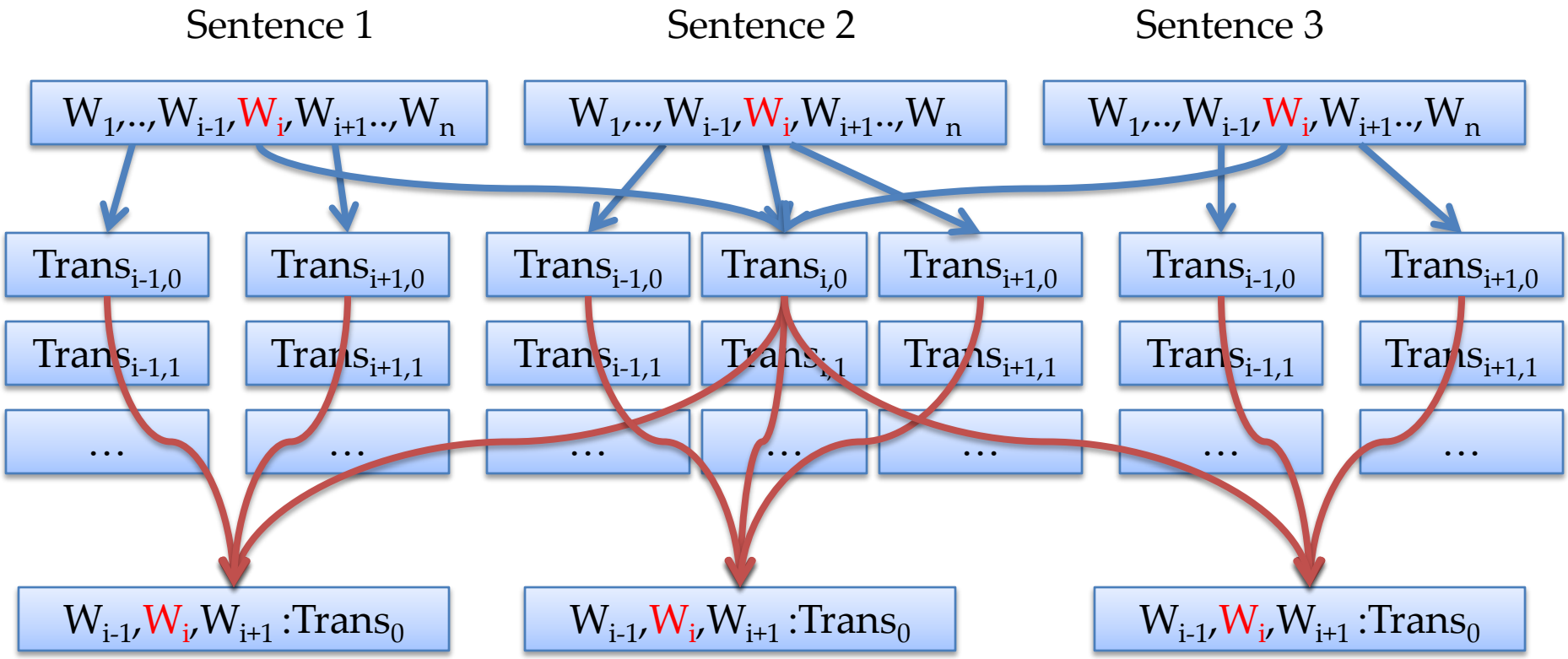
Term Recognition

Running Example



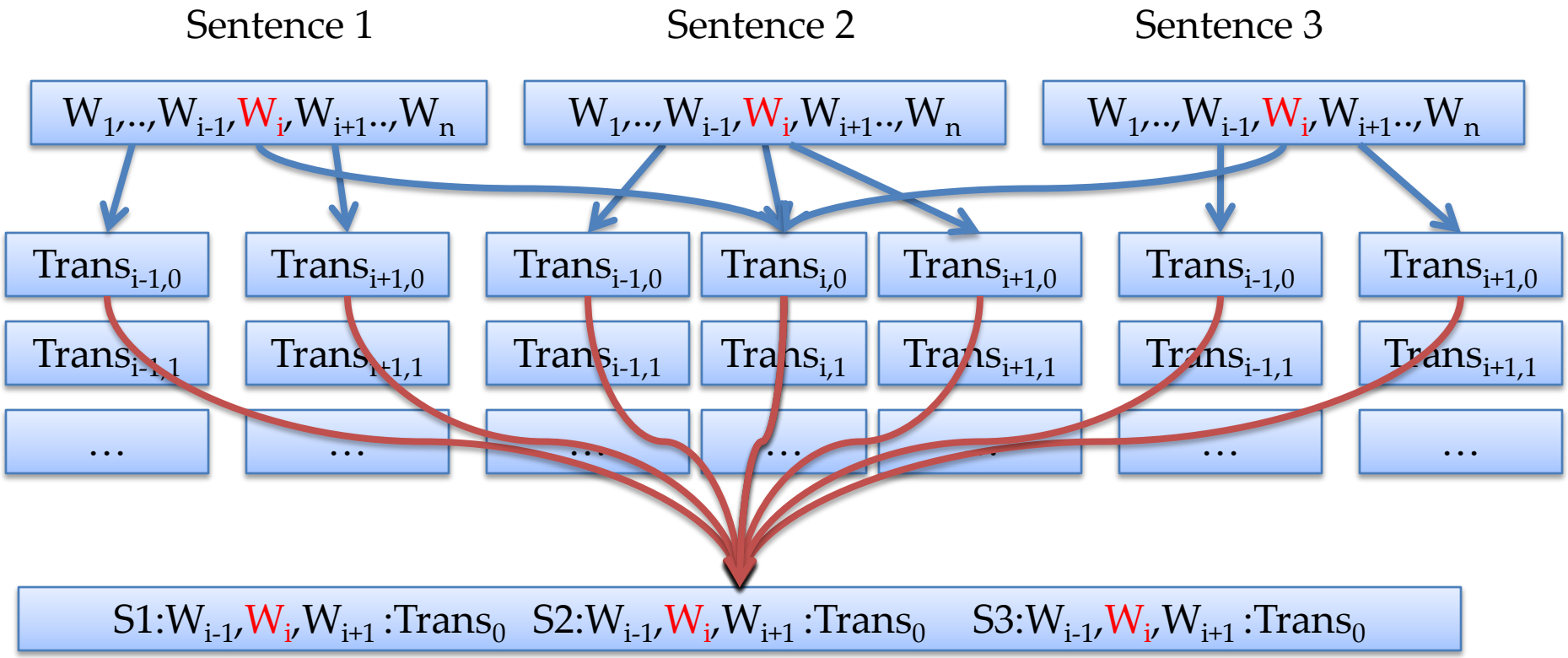
Generating Translation for each Cube
Linking Translation Consistent Cubes

Running Example



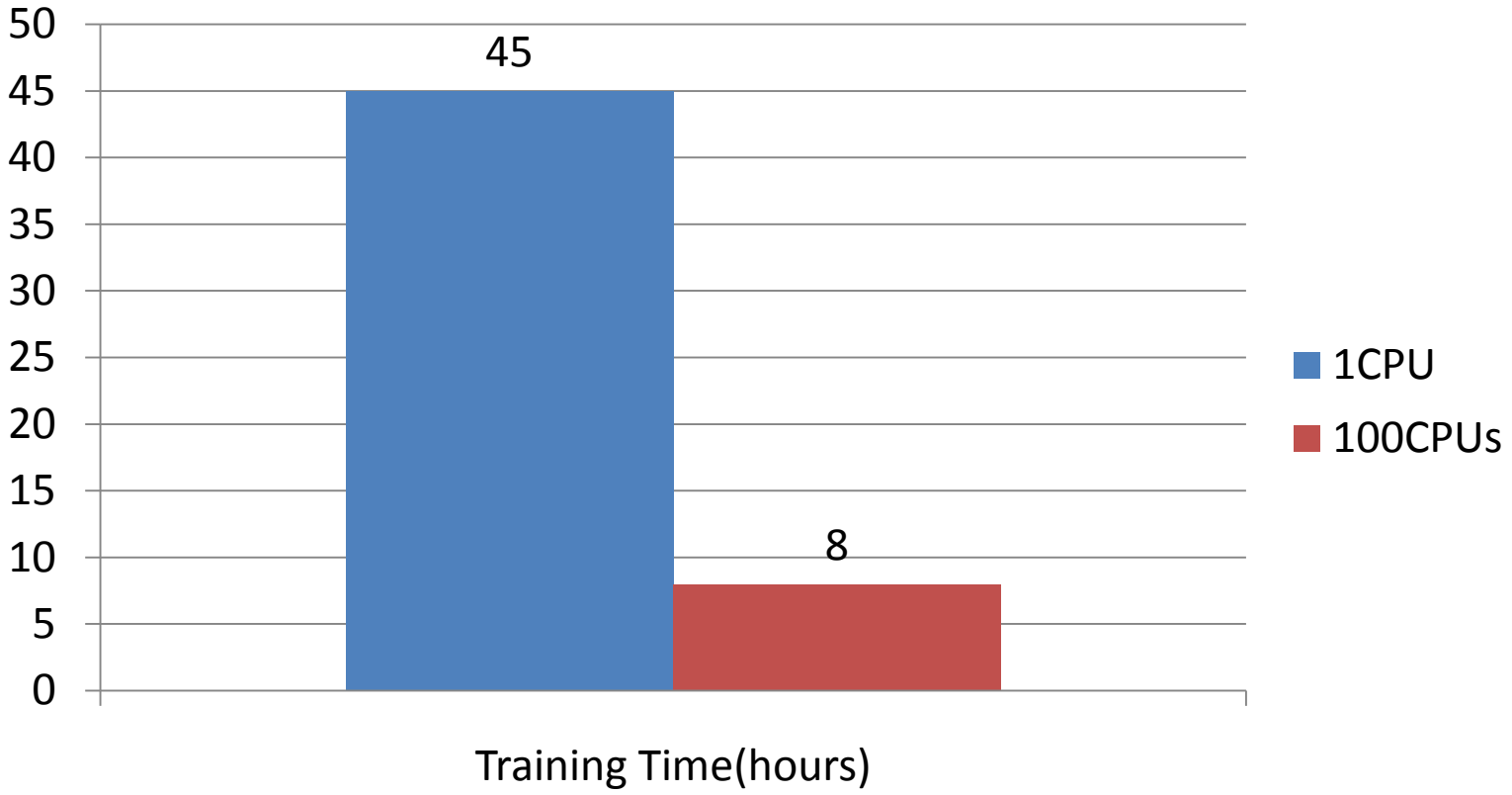
Traditional CKY

Running Example

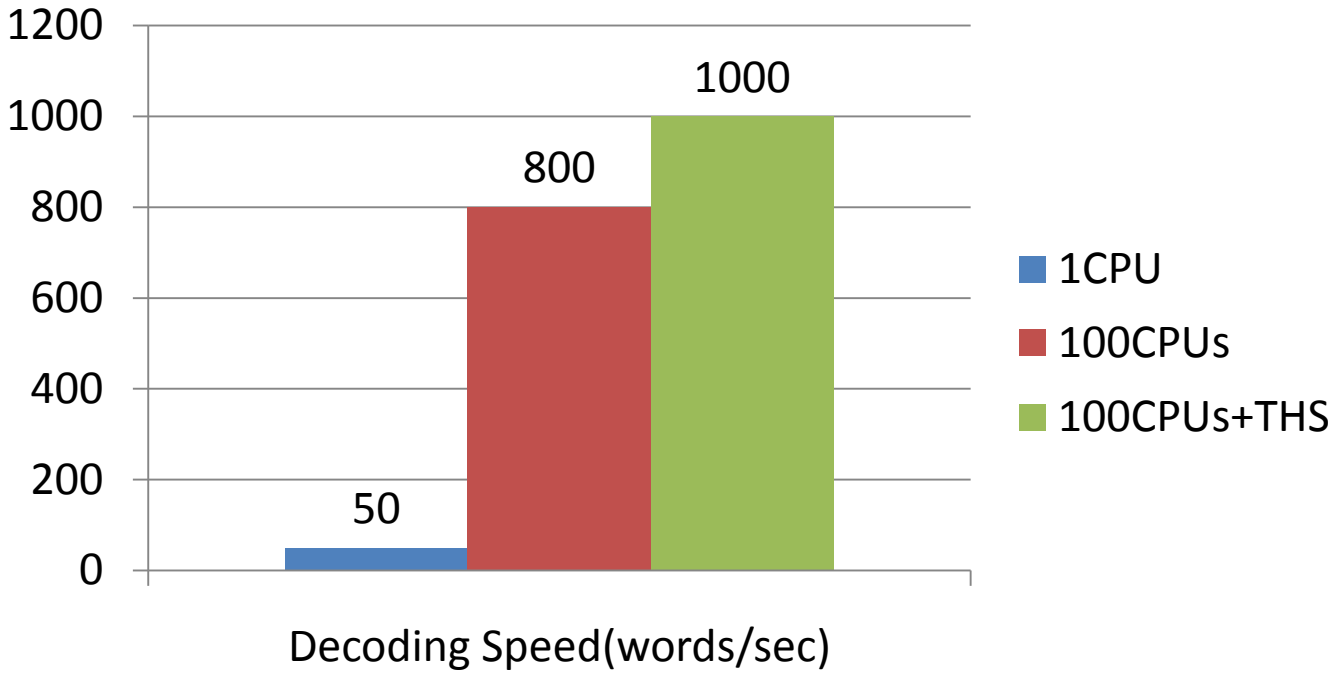


Our CKY: Push into one Cube

Parallel Training



Parallel Decoding



THS: Translation Historical Storage(1 Million Training Data)
Testing Sentences: Extrinsic 10k Sentences

Outline

- Something about TRGTK
- Techniques used in Patent MT
- **Summary**

Summary

- Fast and Stable Translation Service
- Parallel Chinese Segmenter
- Computing on GPUs
- Using Seg/Trans History Storage
- Document-level Decoder

Discussion

- Patent MT helps Translators
  - Google API: reduce 50% time
  - Our API (20M Training Sentences, 1M Terminologies, Refined Segmenter): reduce 90% time

一种治疗肝纤维化及肝硬化的中成药及其制剂的制备方法,属中成药及其制备方法技术领域。其中成药原料组分:郁金、猪苓、赤芍、青皮、木香、大腹皮、泽泻、厚朴(制)、鸡内金、鳖甲。

A Chinese medicinal composition for treating hepatic fibrosis and liver cirrhosis and its preparation process, and belongs to the field of chinese medicine and its prepn technology. Wherein the medicinal material component: Curcmae Rhizoma, pig 苓, Radix Paeoniae Rubra, Pericarpium Citri Reticulatae Viride, Radix Aucklandiae, Pericarpium Arecae, Rhizoma Alismatis, Cortex Magnoliae Officinalis preparata, Endothelium Corneum Gigeriae Galli, Carapax Trionycis.

谢谢！

Thank you！

ありがとう

Danke

Je vous remercie!

감사합니다!

ขอบคุณ!