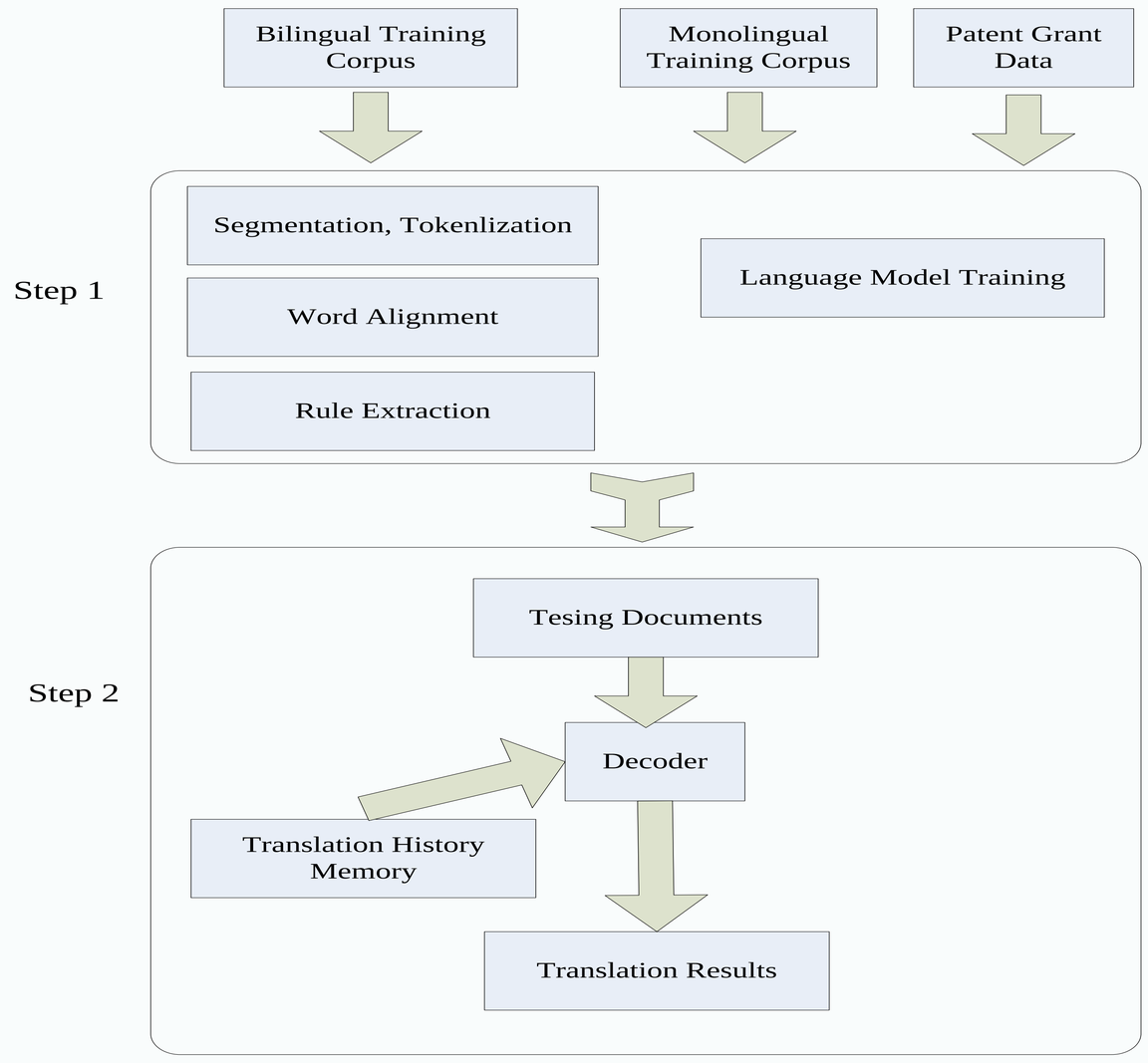
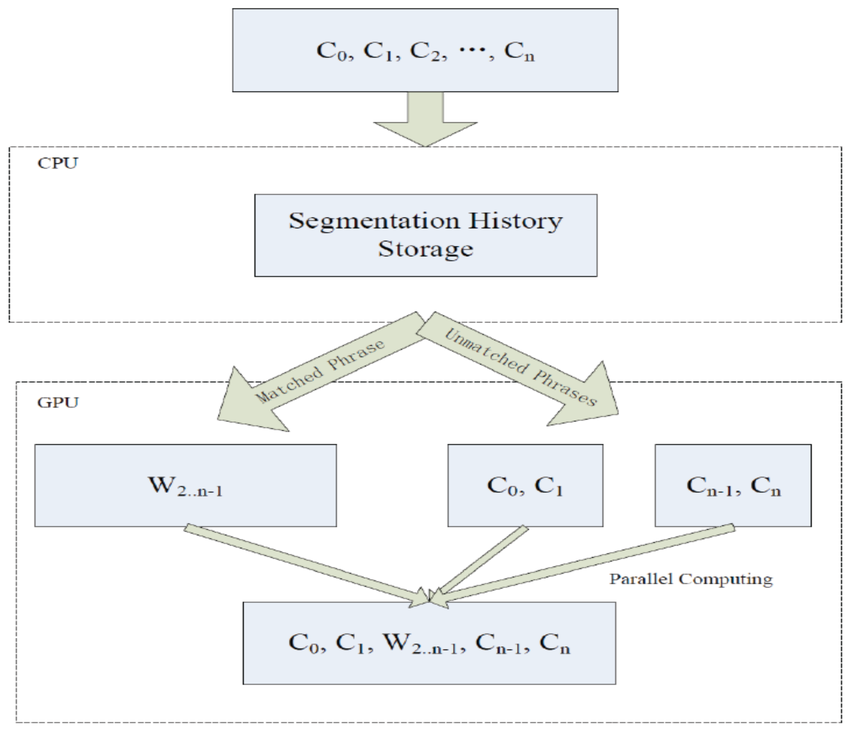
The *TRGTK 's Patent MT System Description for NTCIR-10*

Hao Xiong and Weihua Luo
Torangetek Inc.
Key Lab. of Intelligent Information Processing
{xionghao, luoweihua}torangetek.com

Overall Architecture



Parallel Segmenter



Term Recognition

C-Value

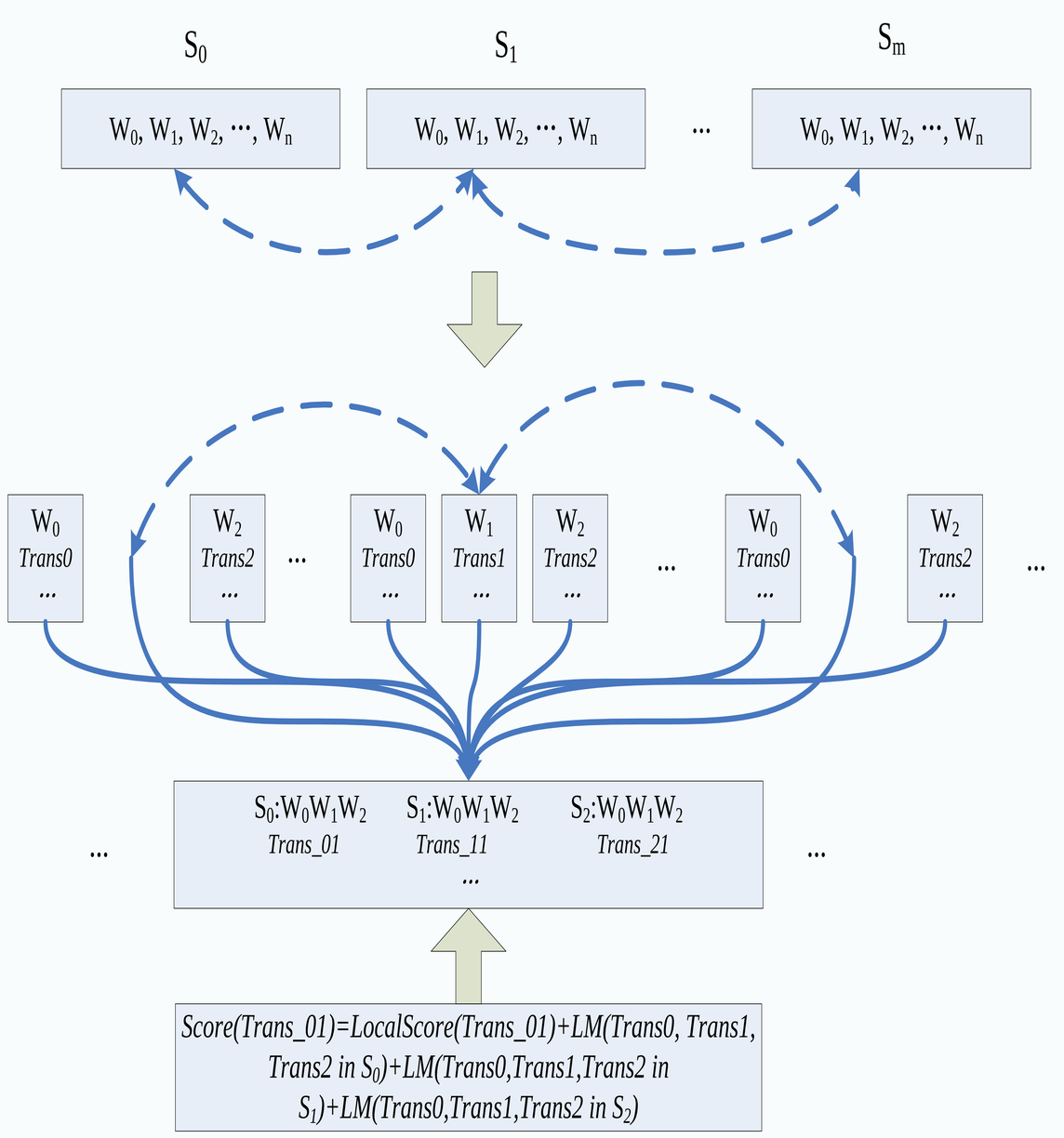
$$C\text{-}value(a) = \begin{cases} log_2|a| \cdot f(a), & a \text{ is not nested,} \\ log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) \end{cases}$$

Candidate Strings

2 words	3 words	4 words	5 words	6 words
n+n	n+n+n	n+n+n+n	v+v+n+n+n	n+n+c+v|n+n+n
n+v	v+n+n	n+n+v+n	d+v+n+n+n	n+n+v|n+c+v|n+n
v+n	n+v+n	v+n+n+n	m+v+m+n+n	n+n+u+b+v|n+n
a+n	v+v+n	v+n+v+n	b+v+n+v+n	v|n+n+v|n+c+v|n+n
d+n	b+v+n	n+v+v+n	n+n+v+n+n	n+v|n+u+n+v|n+n
b+n	n+m+n	v+v+n+n v+n+b+n	a+n+v+n+n	

a is adjective, b is distinguish word, c is conjunction, d is adverb, n is noun, m is numbers, v is verb, u is particle, v|n *is verb or noun.*

Document-level Decoding



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

System	C-E	J-E	E-J
IE	34.63	26.99	32.21
ChE	33.46	26.34	31.4
ME	21.52	26.34	

Parallel Segmenter

Sys1:use segmentation historical storage(SHS)

Sys2: does not use SHS

Corpus: 1 million training data+10 million Chinese sentences generated SHS

System	GPU	Time(minutes)
Sys1	500	0.3
Sys1	1	32
Sys2	500	0.5
Sys2	1	40

Parallel Decoding

System	CPU	Time
Parallel training	100	8 hours
Training	1	45 hours
Parallel decoding	100	800words/s
Parallel decoding+Translation Memory	100	1000words/s
Decoding	1	50words/seconds