Simple and Unsupervised Chinese Word Segmentation

Director: Jungyeul Park

Members: Yilei Zhao
Xingyu YAN

Chinese Word Segmentation

Example:

（C Language is a general computer programming language.）

C 语言是一种通用的计算机程序语言。

Chinese Word Segmentation

Example:

（C | Language | is | a | general | computer |
programming | language.）

C 语言是一种通用的计算机程序语言。

Chinese Word Segmentation

Example:

（C Language is a general computer programming language.）

C 语言是一种通用的计算机程序语言。

{C}{ 语言 **}{** 是 **}{** 一种 **}{** 通用的 **}**

{ 计算机 **}{** 程序 **}{** 语言 **}{** 。 **}**

Chinese Word Segmentation

Example:

{C}{ 语言 **}{** 是 **}{** 一种 **}{** 通用的 **}**

{ 计算机 **}{** 程序 **}{** 语言 **}{** 。 **}**

Good Segmentation
benefits
latter advanced data processing.

Chinese Word Segmentation

Supervised HMM:

Tags: **{S**ingle**,B**egin**,M**iddle**,E**nd**}**

{ 是 /S}

is

{ 计 /B}{ 算 /M}{ 机 /E}

computer

Chinese Word Segmentation

Supervised HMM:

Tags: **{S**ingle**,B**egin**,M**iddle**,E**nd**}**

{C/S}{语/B}{言/E}{是/S}
{一/B}{种/E}{通/B}{用/M}{的/E}

{计/B}{算/M}{机/E}{程/B}{序/E}
{语/B}{言/E}{。/S}

Chinese Word Segmentation

Supervised HMM:

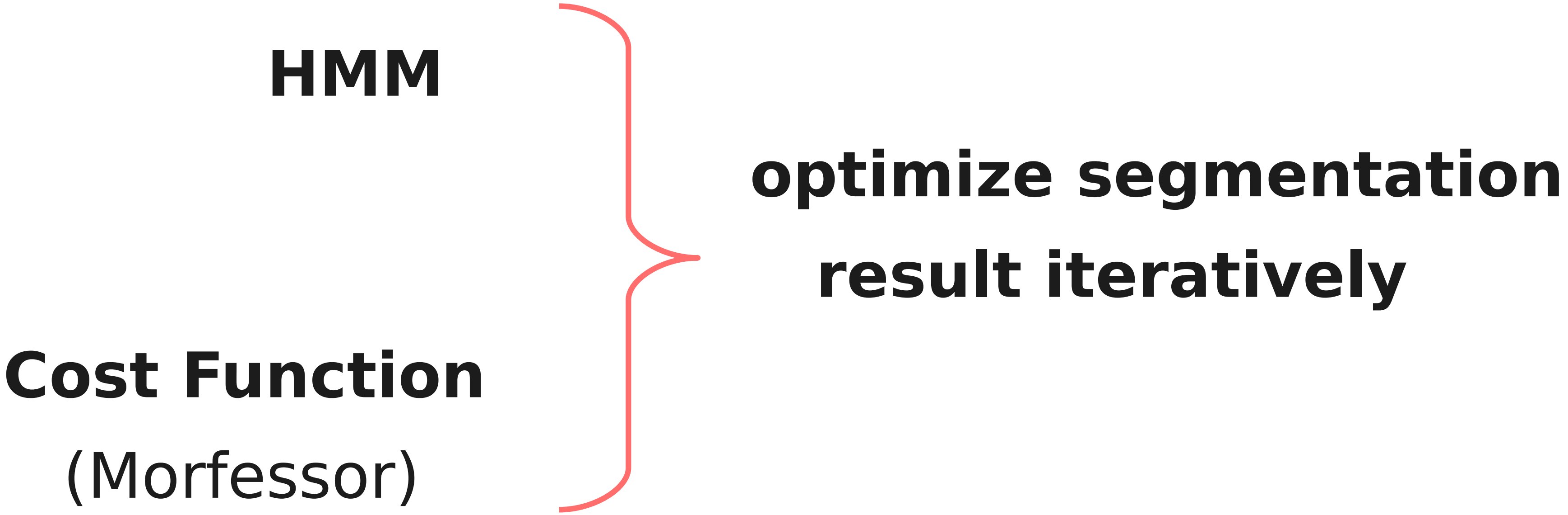
Tags: **{S**ingle**,B**egin**,M**iddle**,E**nd**}**

Easy to achieve 85% accuracy.

Requires tagged training data.

Proposed Solution

Combination of Two Models



HMM

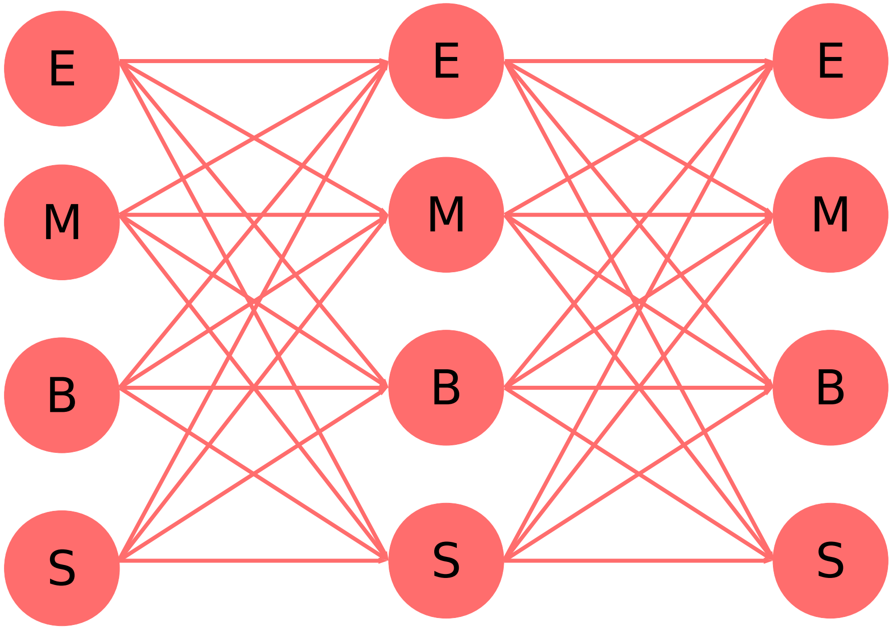
Cost Function

(Morfessor)

optimize segmentation result iteratively

Unsupervised HMM

Character Based Model:



{ 计 /B}　{ 算 /M}　{ 机 /E}

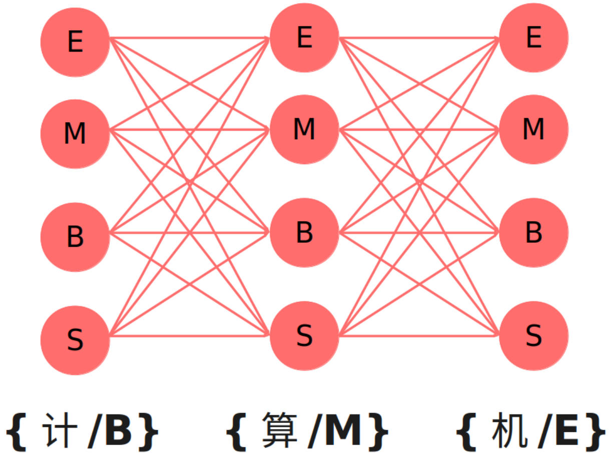
Unsupervised HMM

Issue:

> **without tagged training data, the result is just like guessing (accuracy 50%)**

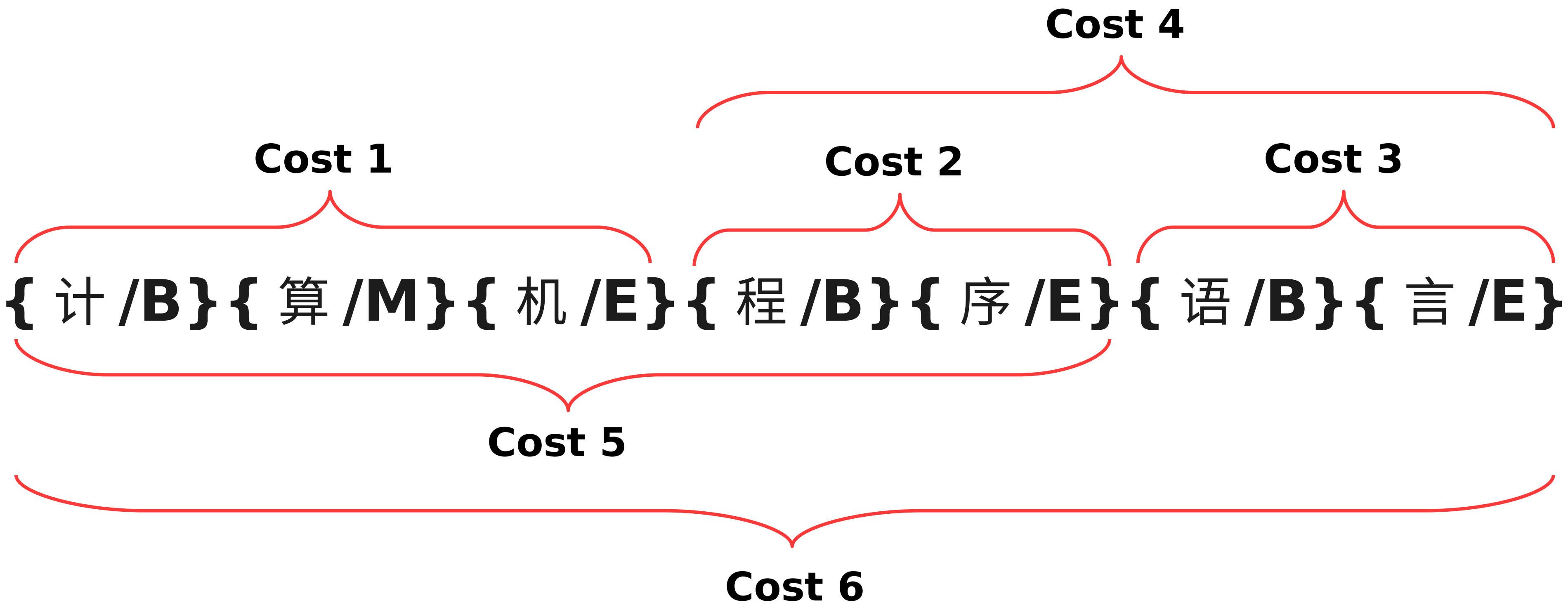
Advantage:

> **inner transition and emission relations**



Cost Function (Morfessor)

Word Based Model:



Cost 4

Cost 1

Cost 2

Cost 3

{计/B}{算/M}{机/E}{程/B}{序/E}{语/B}{言/E}

Cost 5

Cost 6

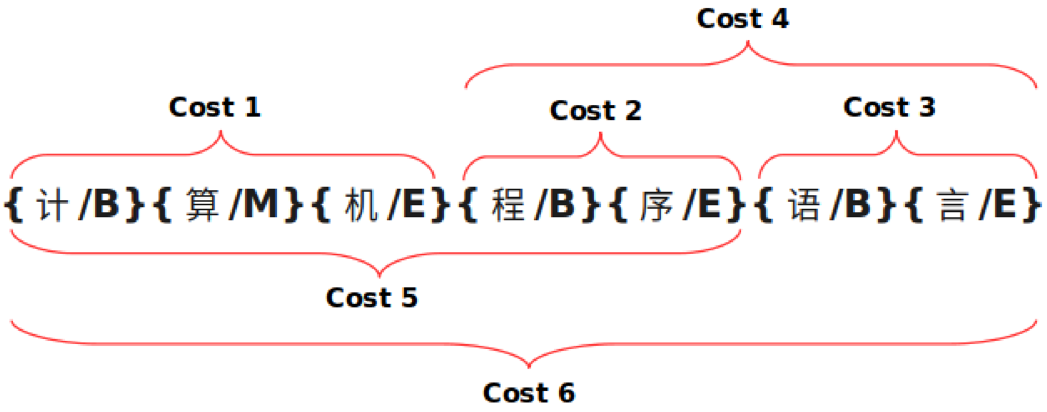
Cost Function (Morfessor)

Issue:

need an initial cookbook (dictionary)

Advantage:

describes the relation between words



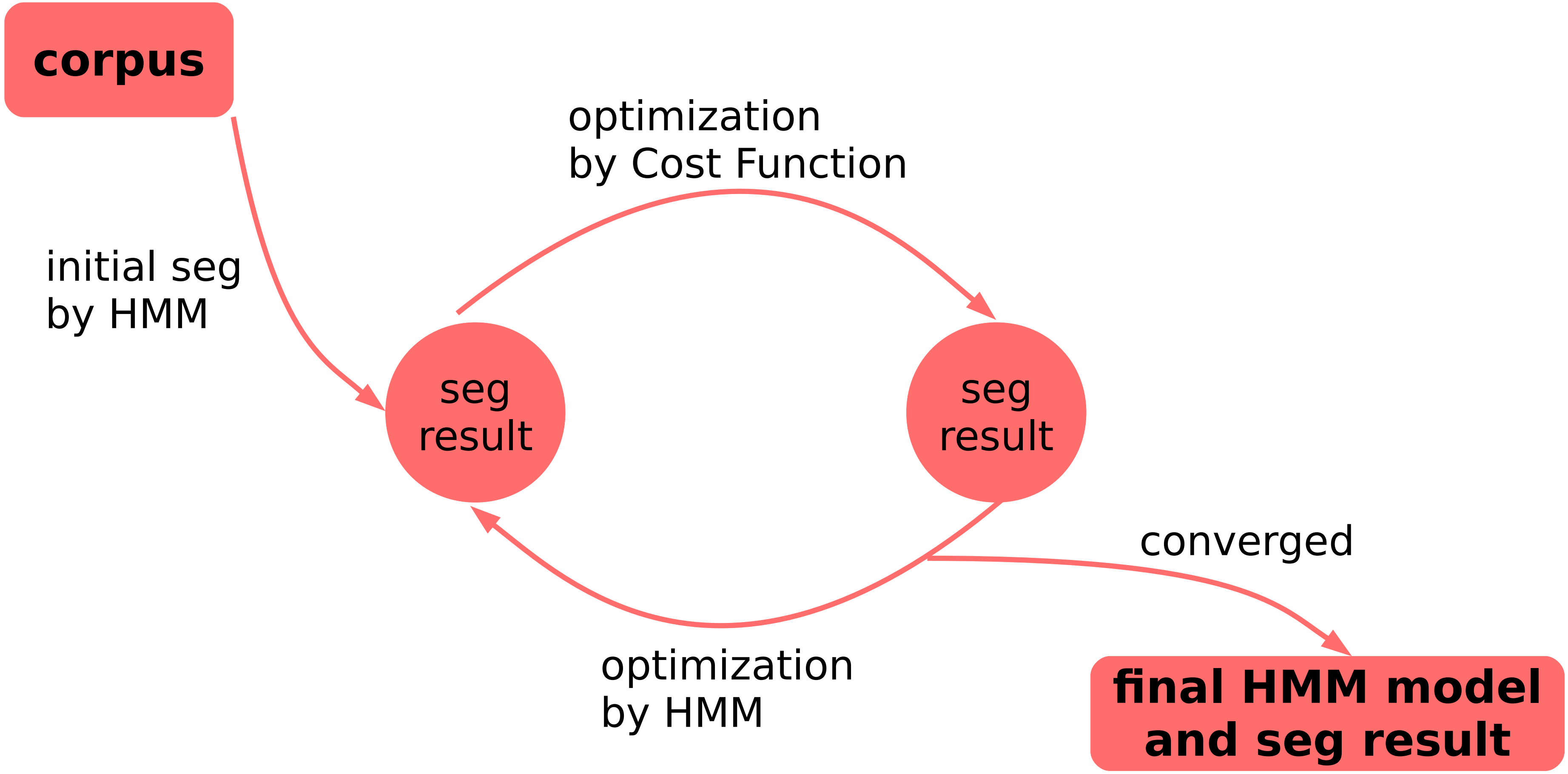
{计/B}{算/M}{机/E}{程/B}{序/E}{语/B}{言/E}

Process

Given un-tagged corpus:

1. segment the corpus by HMM.
2. get the cookbook from HMM result.
3. use Cost Function based method to optimize segmentation result.
4. use result from step 3, get new transition and emission probability. segment the corpus again.
5. repeat step 2–4, until converge.

Process


corpus
initial seg
by HMM
optimization
by Cost Function
seg
result
seg
result
converged
optimization
by HMM
final HMM model
and seg result

Current Works

Current Works

1. **HMM: use NLTK**
2. **Cost Function: not finished yet**
3. **Combination: not finished yet**

Thank you !