Discover Chinese Words from Unsegmented Text

Xianping Ge, Wanda Pratt, Padhraic Smyth
Information and Computer Science
University of California, Irvine
{xge,pratt,smyth}@ics.uci.edu

Problem

- No explicit word boundaries (i.e., spaces) in Chinese text.
- *But*, we need the words for
  - indexing in information retrieval (IR),
  - natural language understanding.

Solution

- Use word probabilities to segment sentences into words.
- Discover the words and their probabilities from raw, unsegmented text using the Expectation-Maximization (EM) algorithm.

Words are running together ...

- No explicit word boundaries (i.e., spaces) in a Chinese sentence. E.g.,
  今天，墨西哥与欧盟就签署自由贸易协定问题
  在此间正式开始谈判。
- The words in the above Chinese sentence:
  今天∧ 墨西哥∧ 与∧ 欧盟∧ 就∧ 签署∧ 自由∧ 贸易∧ 协定∧ 问题
  ∧ 在∧ 此间∧ 正式∧ 开始∧ 谈判
- Compare the English sentence WITHOUT / WITH the spaces between words:

  "*Today,theEuropeanUnion(EU)andMexicostartednegotiations toestablishazoneoffreetradebetweenthem.*"

  "Today, the European Union (EU) and Mexico started negotiations to establish a zone of free trade between them."

Segment the sentence into words: the maximum-likelihood approach

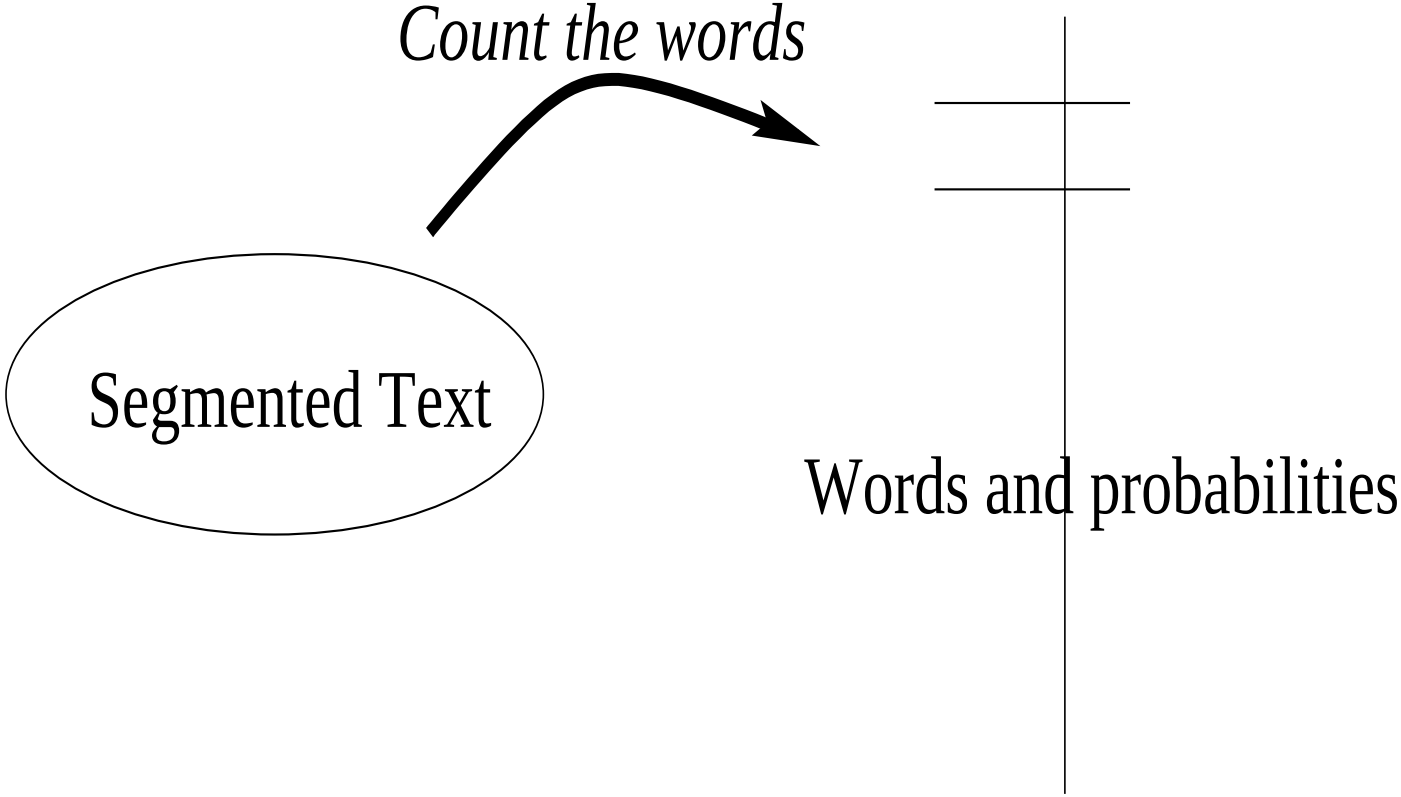
- Segment the sentence $C_1C_2C_3\ldots C_n$ into words $W_1W_2\ldots W_m$ to maximize the likelihood $P(W_1)P(W_2)\ldots P(W_m)$.

	Segmentation	likelihood
	$C_1 \wedge C_2 \wedge C_3$	0.01
$\surd$	$C_1C_2 \wedge C_3$	<u>0.09</u>
	$C_1 \wedge C_2C_3$	0.001
	$C_1C_2C_3$	0.03

- This can be done by dynamic programming *if we knew the words* $\{W_i\}$ *and their probabilities* $P(W_i)$.

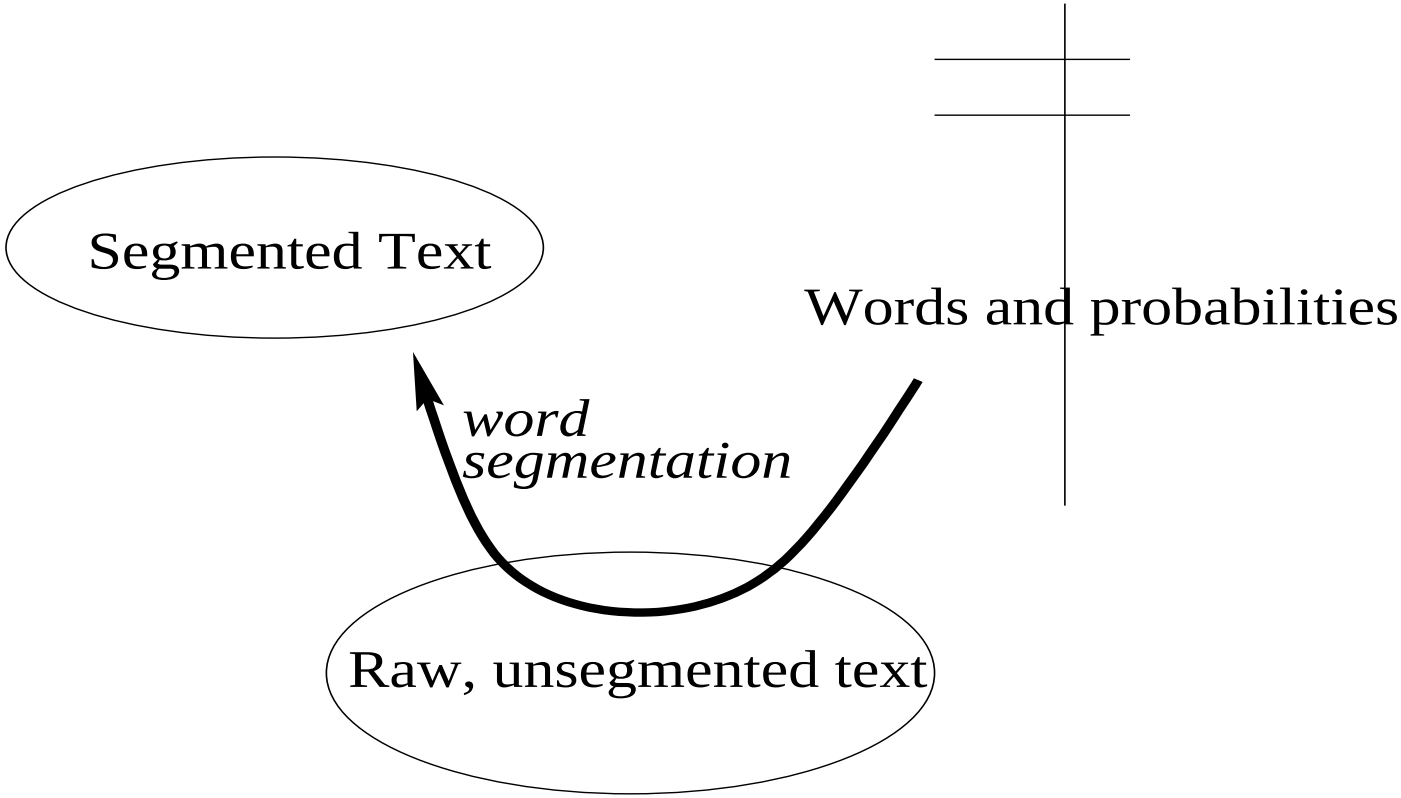
If we had a training corpus of segmented text ...

- then we can easily get the words and their probabilities!



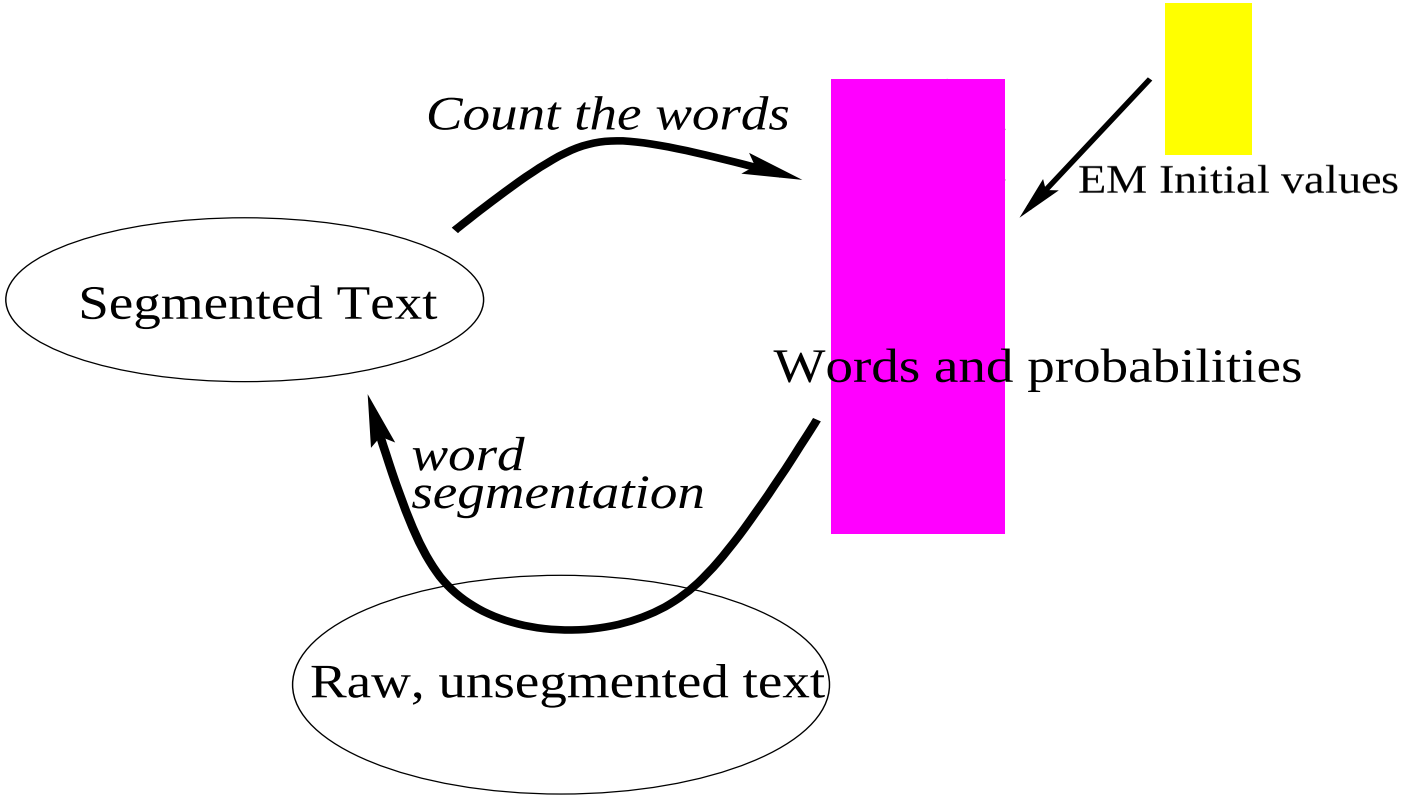
If we knew the words and their probabilities ...

- then we can get the segmented text!



Combine them together: Expectation-Maximization (EM) algorithm

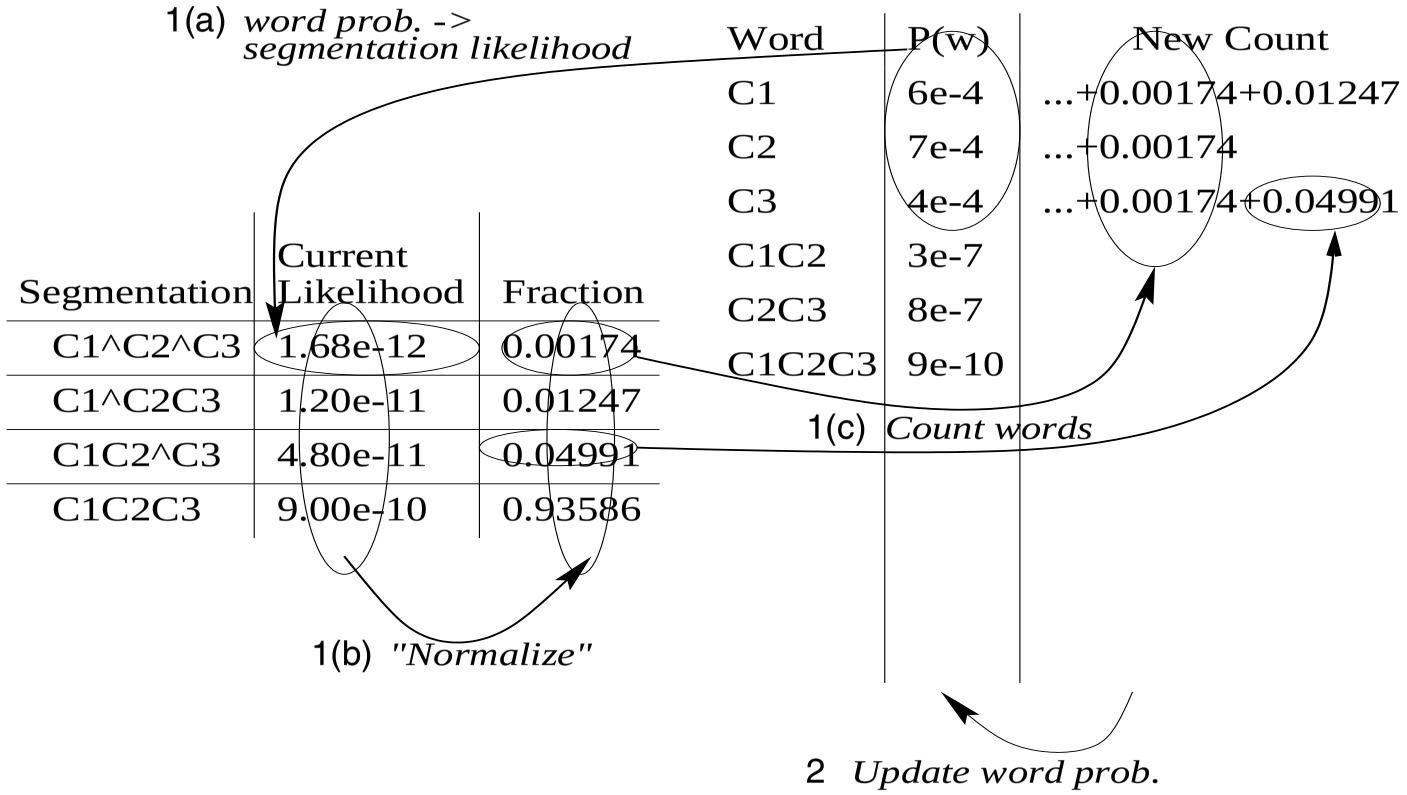
- Solves this problem of "*Which came first, the chicken or the egg*?" by providing the first "egg", i.e., the initial values for the word probabilities.



The main procedure

1. For each sentence in the unsegmented text,
   (a) Compute the likelihood of each possible segmentation using the current estimated values of the word probabilities.
   (b) The segmentation likelihood is normalized as "fraction" that sums to 1.
   (c) Count the words in each segmentation. I.e., add the "fraction" of the segmentation to the word count.
2. Update the word probabilities using the word counts.
3. Repeat until convergence.

An example


1(a) word prob. -> segmentation likelihood
Word
P(w)
New Count
C1
6e-4
...+0.00174+0.01247
C2
7e-4
...+0.00174
C3
4e-4
...+0.00174+0.04991
C1C2
3e-7
C2C3
8e-7
C1C2C3
9e-10
Segmentation
Current Likelihood
Fraction
C1^C2^C3
1.68e-12
0.00174
C1^C2C3
1.20e-11
0.01247
C1C2^C3
4.80e-11
0.04991
C1C2C3
9.00e-10
0.93586
1(b) "Normalize"
1(c) Count words
2 Update word prob.

It works!

- The algorithm correctly discovers most words from the unsegmented text.
  - 但∧ 艺术家∧ 要尊重∧ 艺术∧ 规律
  - 城市∧ 是∧ 先进∧ 生产力∧ 的∧ 集聚∧ 地∧ 和∧ 辐射源
  - 香港∧ 特区∧ 是∧ 祖国∧ 的∧ 一个∧ 组成∧ 部分
  - 信用社∧ 体制∧ 改革和∧ 业务∧ 发展
- Recall/Precision=65.65%/71.91%
- After splitting single-character "stop" words from content words, Recall/Precision can be boosted up to 97.72%/91.05%

Future work

- Some single-character stop words (like “and”, “of”, “should”, etc.) tend to cling to content words.

要尊重	要∧尊重
shouldrespect	should respect

  A more principled way of splitting these stop words from content words?

- Incorporate prior knowledge:
  - Distribution of word lengths
  - Existing word lists
  - Part of speech information
- Other applications

References

[1] Aitao Chen et al. Chinese text retrieval without using a dictionary. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, 1997.

[2] R. Sproat and C. Shih. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4:336–351, March 1990.

[3] Jay M. Ponte and W. Bruce Croft. Useg: A retargetable word segmentation procedure for information retrieval. In *Symposium on Document Analysis and Information Retrieval 96 (SDAIR)*, 1996.

[4] Q. An and W. S. Wong. Automatic segmentation and tagging of Hanzi text using a hybrid algorithm. In *Proceedings of the 9th International Conference on Industrial & Engineering Applications of AI & Expert Systems*, June 4-7 1996.

[5] M. R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. Machine Learning 34, 1-3, February 1999.