Unsupervised Tokenization for Machine Translation

Author: Tagyoung Chung et al.

Presenter: Jae Dong Kim

April 14, 2010

Introduction

- Observations
  - Chinese
    - no space
  - Korean, Hungarian
    - with space but coarse granularity
      - a single word consists of multiple morphemes → corresponds to separate words in English
      - compound words
  - Tokenization for MT
    - the first step
    - one to one mapping of words will be ideal
    - gold standard tokenization does not necessarily help MT
    - statistical methods require hand-annotated data

Introduction

- This work
  - unsupervised methods to find an appropriate tokenization for MT
  - method using parallel data vs. method using monolingual data
  - Chinese
    - no space
  - Korean
    - smaller-scale tokenization

Tokenization

- isolating languages
  - English, Chinese
  - one word equals a single token
- agglutinative languages
  - Hungarian, Japanese, Korean
  - token boundary is ambiguous
    - szekrenyemben (in my closet : closet-my-inessive) → szekreny em ben
    - meok-eoss-da (ate : eat-past-indicative) → meok eoss da ??

Model 1

- Learning Tokenization from alignment
- Input : English words $\mathbf{e}_1^n$, Foreign characters $\mathbf{c}_1^m$
- Unobserved variables: word-level alignments, tokenizations
- tokenization with a string: $\mathbf{s}_1^m$
- string of foreign words: $\mathbf{f}_1^l$
- Using IBM model 1 → P(a) = 1/n, but P(f|e) ?

$$\mathrm{f} = \mathrm{s} \circ \mathrm{c} \quad \text{where} \quad \ell = \sum_{i=1}^{m} s_i$$

$$\begin{aligned} P(\mathrm{f}|\mathrm{e}) &= \sum_{\mathrm{a}} P(\mathrm{f}, \mathrm{a} \mid \mathrm{e}) \\ &= \sum_{\mathrm{a}} \prod_{i} P(f_i \mid e_{a_i}) P(\mathrm{a}) \\ &= \prod_{i} \sum_{j} P(f_i \mid e_j) P(a_i = j) \end{aligned}$$

Model 1

- posterior prob. of a word beginning at position i, ending at position j, and being generated by English word k:

$$P(\mathrm{s}_{i\ldots j} = (1,0,\ldots,0,1), a = k \mid \mathrm{e}) = \frac{\alpha(i)P(f \mid e_k)P(a = k)\beta(j)}{P(\mathrm{c} \mid \mathrm{e})}$$

$$\alpha(i) = P(\mathrm{c}_1^i, s_i = 1 \mid \mathrm{e})$$

$$\beta(j) = P(\mathrm{c}_{j+1}^m, s_j = 1 \mid \mathrm{e})$$

$$\alpha(i) = \sum_{\ell=1}^{L} \alpha(i-\ell) \sum_a P(a)P(c_{i-\ell}^i \mid e_a)$$

$$\beta(j) = \sum_{\ell=1}^{L} \sum_a P(a)P(c_j^{j+\ell} \mid e_a)\beta(j+\ell)$$

$$ec(c_i^j, e_k) \mathrel{+}= \frac{\alpha(i)P(a)P(c_i^j \mid e_k)\beta(j)}{\alpha(m)}$$

The M step simply normalizes the counts:

$$\tilde{P}(f \mid e) = \frac{ec(f,e)}{\sum_e ec(f,e)}$$

Model 1

- a* → s* → optimal segmentation of f

$$\mathbf{a}^* = \operatorname*{argmax}_{\mathbf{a}} P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

- vs. HMM
  - a target word generates a source token
  - transition → segmentation
  - emission → alignment
  - HMM-like dynamic programming to do inference

Model 2

- Monolingual
  - P(s) are equally likely

$$P(\mathrm{s} \mid \mathrm{c}) \propto P(\mathrm{c} \mid \mathrm{s}) P(\mathrm{s})$$

$$P(\mathrm{f}) = P(f_1) \times \ldots \times P(f_\ell)$$

$$P(f_i) = \frac{count(f_i)}{\sum_k count(f_k)}$$

$$\mathrm{s}^* = \operatorname*{argmax}_{\mathrm{s}} P(\mathrm{f})$$

New data

- sentence to be translated is monolingual

$$P(f) = \sum_{e} P(f \mid e) P(e)$$

Preventing Overfitting

- Variational Bayes

$$\theta_e \mid \alpha \sim \mathrm{Dir}(\alpha),$$
$$f_i \mid e_i = e \sim \mathrm{Multi}(\theta_e).$$

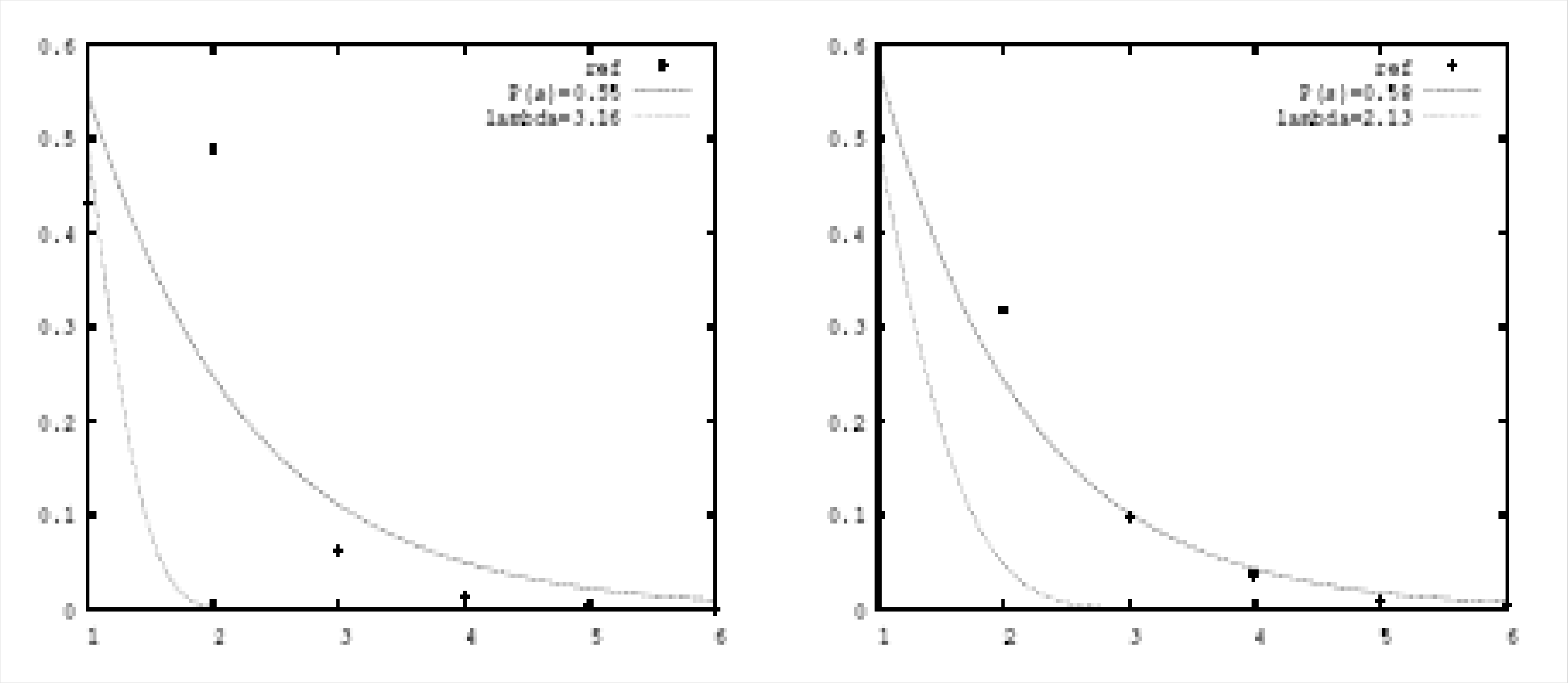
$$\tilde{P}(f \mid e) = \frac{\exp(\psi(ec(f,e) + \alpha))}{\exp(\psi(\sum_e ec(f,e) + s\alpha))}$$

Preventing Overfitting

- Token Length

$$\phi_1(\ell) = P(s)(1 - P(s))^{\ell - 1}$$

$$\phi_2(\ell) = 2^{-\ell^{\lambda}}$$



Preventing Overfitting

- Token Length
  - Model 2

$$P(\mathrm{f}) \propto P(f_1)\phi(\ell_1) \times \ldots \times P(f_n)\phi(\ell_n)$$

  - Model 1

$$\alpha(i) = \sum_{\ell=1}^{L} \alpha(i-l)\phi_1(\ell) \sum_a P(a)P(c_{i-\ell}^{i} \mid e_a)$$

and the expected count of (c_i^j, e_k) is

$$ec(c_i^j, e_k) \mathrel{+}= \frac{\alpha(i)P(a)P(c_i^j \mid e_k)\beta(j)\phi_1(j-i)}{\alpha(m)}$$

$$P(s) = \frac{1}{m}\sum_{i}^{m} \frac{\alpha(i)\beta(i)}{\alpha(m)}$$

Data

- Chinese-English
  - FBIS newswire data
  - Dev, Test: 1,000 with 10 refs. each
- Korean-English
  - collected from news websites
  - Training set: 60K - 2,200
  - Dev, Test: 1,100 with 1 ref. each

Experimental Setup

- Moses
  - GIZA++ was run until the perplexity on dev set stopped decreasing
- Maximum size of a token (L)
  - 3 for Chinese, 4 for Korean
- Compared to supervised segmenters
  - Chinese: LDC, Xue's, PKU & CTB
  - Korean: Rule-based Morphological Analyzer
-

Results

	Chinese		Korean
	BLEU	F-score	BLEU
Supervised			
Rule-based morphological analyzer			7.27
LDC segmenter	20.03	0.94	
Xue's segmenter	23.02	0.96	
Stanford segmenter (pku)	21.69	0.96	
Stanford segmenter (ctb)	22.45	1.00	
Unsupervised			
Splitting punctuation only			6.04
Maximal (Character-based MT)	20.32	0.75	
Bilingual $P(f \mid e)$ with ϕ_1 $P(s) = learned$	19.25		6.93
Bilingual $P(f)$ with ϕ_1 $P(s) = learned$	20.04	0.80	7.06
Bilingual $P(f)$ with ϕ_1 $P(s) = 0.9$	20.75	0.87	**7.46**
Bilingual $P(f)$ with ϕ_1 $P(s) = 0.7$	20.59	0.81	7.31
Bilingual $P(f)$ with ϕ_1 $P(s) = 0.5$	19.68	0.80	7.18
Bilingual $P(f)$ with ϕ_1 $P(s) = 0.3$	20.02	0.79	7.38
Bilingual $P(f)$ with ϕ_2	**22.31**	0.88	7.35
Monolingual $P(f)$ with ϕ_1	20.93	0.83	6.76
Monolingual $P(f)$ with ϕ_2	20.72	0.85	7.02

Results

- performance with $p(f|e)$ < performance with $p(f)$
  - consistency is important
- bilingual is better
  - learning boundaries from the target language
- the second length factor was better
  - need for heavy discount for longer tokens
- higher F → higher BLEU?

Future Work

- applied to one language of the pair
  - one isolating, one synthetic
  - could be extended to tokenization for both languages.
- the most simple alignment model
  - more complex model

Discussion

- Does the basic model 1 encourage 1 to 1 mapping?
- De-segmentation for MT performance?
  - ex) segmentation for alignment, and then, de-segmentation on English

Korean Tokenization

English	the two presidents will hold a joint press conference at the end of their summit talks .
Untokenized Korean	양국 정상은 회담이 끝난뒤 공동 기자회견을 갖고 회담 결과를 공식 발표한다 .
Supervised	양국 정상␣은 회담␣이 끝나␣ㄴ 뒤 공동 기자회견␣을 갖␣고 회담 결과␣를 공식 발표␣하␣ㄴ다 .
Bilingual $P(f \mid e)$ with ϕ_1	양국 정상은 회담␣이 끝난 뒤 공동 기자회견␣을 갖고 회담 결과␣를 공식 발표한␣다 .
Bilingual $P(f)$ with ϕ_2	양국 정상␣은 회담␣이 끝난 뒤 공동 기자회견␣을 갖고 회담 결과␣를 공식 발표␣한다 .
Monolingual $P(f)$ with ϕ_1	양국 정␣상␣은 회담␣이 끝난 뒤 공동 기자회견␣을 갖고 회담 결과를 공식 발표한␣다 .
Monolingual $P(f)$ with ϕ_2	양국 정상␣은 회담␣이 끝난 뒤 공동 기자회견␣을 갖고 회담 결과를 공식 발표␣한다 .