

Ruey-Cheng Chen <rueycheng@turing.csie.ntu.edu.tw>

Regularized Compression (Chen et al., 2012)

The mechanism is analogous to digram coding **Notation** (Witten et al., 1999). One starts with a sequence of single-character words (W_0) and works from that representation up in an agglomerative fashion, iteratively removing boundaries between two selected word types (effectively producing W_i from W_{i-1} .) Regularized compression employs a specialized decision criterion for balancing compression rate and vocabulary complexity:

 $-\alpha f(x, y) + |W_{i-1}| \Delta H(W_{i-1}, W_i)$ min. x and y are word types; $f(x, y) > n_{ms}$; s.t. Note that we still need to estimate α and n_{ms} . either x or y is a character.

Change in Description Length

Compression Moves Masses Let x and y de- **Approximation** The second term in the original note the selected word types. Let z = xy be a objective can be approximated by the change in new unseen token introduced to replace all the description length between W_i and W_{i-1} : bigrams (x, y). The following summarizes the change in observed frequencies.

	f(x)	f(y)	f(z)	W
W_{i-1}	k	l	0	N
W_i	k-m	l-m	m	N-m

Analysis Note that the first three lines in the last equation are of the form $x_1 \log x_1 - x_2 \log x_2$ for some $x_1, x_2 \ge 0$. By using the Taylor series, we have the following inequalities:

$$m\log\frac{(k-m)(l-m)}{Nm} \le \Delta L \le m\log\frac{kl}{(N-m)m}.$$
(1)

New Objectives

• G_1 : Replacing the second term in the original objective with the *lower bound*.

$$f(x,y) \left(\log \frac{(f(x) - f(x,y))(f(y) - f(x,y))}{|W_{i-1}|f(x,y)} - \alpha \right)$$

• G_2 : Same as G_1 except that the lower bound is divided by f(x, y) beforehand.

$$-\alpha f(x,y) + \log \frac{(f(x))}{1}$$

An Improved MDL-Based Compression Algorithm for Unsupervised Word Segmentation

- f(x, y): bigram frequency;
- $|W_{i-1}|$: sequence length of W_{i-1} ;
- $\Delta H(W_{i-1}, W_i) = H(W_i) H(W_{i-1}).$

We estimate the Shannon entropy H(W) empirically using maximum likelihood, as in:

$$\log|W| - \frac{1}{|W|} \sum_{x:types} f(x) \log f(x).$$

$$\Delta L = [(N - m) \log(N - m) - N \log N]$$

+ $[k \log k - (k - m) \log(k - m)]$
+ $[l \log l - (l - m) \log(l - m)]$
- $m \log m$

$$\frac{f(x,y)(f(y) - f(x,y))}{|W_{i-1}|f(x,y)|}$$

(in the sense of description length); (b) Find α as in (a). Fix α and vary ρ ; (c) Set ρ to a heuristic value 0.37 and vary α . We use the following procedure to compute description length (Rissanen, 1978). Given a word sequence W (say M types in total), we write out all the induced word types entry by entry as a character sequence C. Then the overall description length is:

Performance Chart Method Adaptors grammar, colloc3-syllable Johnson and C Regularized compression/MDL, $G_2(b)$ Regularized compression/MDL Chen et al. (20 Adaptors grammar, colloc Johnson and (Particle filter, unigram Börschinger ar Regularized compression/MDL, $G_1(b)$ Bootstrap voting experts/MDL Hewlett and C Nested Pitman-Yor process, bigram Mochihashi et Branching entropy/MDL Zhikov et al. (Börschinger ar Particle filter, bigram Hierarchical Dirichlet process Goldwater et a

Chen, R.-C., Tsai, C.-M., and Hsiang, J. (2012). A regularized compression method to unsupervised word segmentation. In Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, SIGMORPHON '12, pages 26–34, Montreal, Canada. Association for Computational Linguistics. Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. Cognition, 112(1):21-54.

National Taiwan University

Result on Bernstein-Ratner Corpus

Setup Set $n_{ms} = 3$ as suggested. Employ three **Performance** specialized MDL-based search runs for α and ρ (analogous to one-round coordinate ascent):

- (a) Fix ρ to 0 and vary α to find the best value

$$|W|\tilde{H}(W) + |C|\tilde{H}(C) + \frac{M-1}{2}\log|W|.$$

^aTested on an Intel Xeon 2.5GHz 8-core machine with 8GB RAM.

References

Johnson, M. and Goldwater, S. (2009). Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, pages 317–325, Boulder, Colorado. Association for Computational Linguistics.

Run		Ρ	R	F
Baseline		76.9	81.6	79.2
G_1 (a)	lpha:0.030	76.4	79.9	78.1
G_1 (b)	ho: 0.38	73.4	80.2	76.8
G_1 (c)	lpha:0.010	75.7	80.4	78.0
G_2 (a)	lpha:0.002	<u>82.1</u>	80.0	81.0
G_2 (b)	ho: 0.36	79.1	81.7	80.4
G_2 (c)	$\alpha: 0.004$	79.3	<u>84.2</u>	<u>81.7</u>

Meth Adapt Adapt Regul

Regul Regul



Average Running Time (Per Fold)^a

od	Sec.
tors grammar, colloc3-syllable	53826
tors grammar, colloc	10498
larized compressor	1.51
larized compressor, G_1 (b)	0.60
larized compressor, G_2 (b)	1.25

	Ρ	R	F
Goldwater (2009)	86.1	88.4	87.2
	79.1	81.7	80.4
012)	76.9	81.6	79.2
Goldwater (2009)	78.4	75.7	77.1
nd Johnson (2012)	—	—	77.1
	73.4	80.2	76.8
Cohen (2011)	79.3	73.4	76.2
al. (2009)	74.8	76.7	75.7
(2010)	76.3	74.5	75.4
nd Johnson (2012)	—	—	74.5
al. (2009)	75.2	69.6	72.3