An Improved MDL-Based Compression Algorithm for Unsupervised Word Segmentation

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

Regularized Compression (Chen et al., 2012)

The mechanism is analogous to digram coding (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*:

min. $-\alpha f(x,y) + |W_{i-1}|\Delta\tilde{H}(W_{i-1}, W_i)$
s.t. x and y are word types; $f(x,y) > n_{\text{ms}}$;
either x or y is a character.

Notation

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

We estimate the Shannon entropy $\tilde{H}(W)$ empirically using maximum likelihood, as in:

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

Note that we still need to estimate α and n_{ms}.

Change in Description Length

Compression Moves Masses Let x and y denote the selected word types. Let $z = xy$ be a new unseen token introduced to replace all the 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	$N-m$

Approximation The second term in the original objective can be approximated by the change in description length between W_i and W_{i-1}:

$$\begin{aligned}\Delta L = &[(N-m)\log(N-m) - N\log N] \\ &+ [k\log k - (k-m)\log(k-m)] \\ &+ [l\log l - (l-m)\log(l-m)] \\ &- m\log m\end{aligned}$$

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 $x_1, x_2 \geq 0$. By using the Taylor series, we have the following inequalities:

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

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.

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

Result on Bernstein-Ratner Corpus

Setup Set $n_{\text{ms}} = 3$ as suggested. Employ three 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 (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:

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

Performance

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

Average Running Time (Per Fold)[a]

Method	Sec.
Adaptors grammar, colloc3-syllable	53826
Adaptors grammar, colloc	10498
Regularized compressor	1.51
Regularized compressor, G_1 (b)	0.60
Regularized compressor, G_2 (b)	1.25

Performance Chart

Method		P	R	F
Adaptors grammar, colloc3-syllable	Johnson and Goldwater (2009)	86.1	88.4	87.2
Regularized compression/MDL, G_2(b)	—	79.1	81.7	80.4
Regularized compression/MDL	Chen et al. (2012)	76.9	81.6	79.2
Adaptors grammar, colloc	Johnson and Goldwater (2009)	78.4	75.7	77.1
Particle filter, unigram	Börschinger and Johnson (2012)	–	–	77.1
Regularized compression/MDL, G_1(b)	—	73.4	80.2	76.8
Bootstrap voting experts/MDL	Hewlett and Cohen (2011)	79.3	73.4	76.2
Nested Pitman-Yor process, bigram	Mochihashi et al. (2009)	74.8	76.7	75.7
Branching entropy/MDL	Zhikov et al. (2010)	76.3	74.5	75.4
Particle filter, bigram	Börschinger and Johnson (2012)	–	–	74.5
Hierarchical Dirichlet process	Goldwater et al. (2009)	75.2	69.6	72.3

[a] Tested on an Intel Xeon 2.5GHz 8-core machine with 8GB RAM.

References

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.

Johnson, M. and Goldwater, S. (2009). Improving nonparametric 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.