Reading: Decipherment Foreign Language

Yu Usami

March 4, 2013

Noisy Channel Model

"Really written in English, but has been coded in some strange symbols." – [Weaver 1955]

Goal

Translation system from foreign language to English

$$\arg\max_e P(e|f) = \arg\max_e p(e)p(f|e)$$

IBM Model 2

Model $P(f|e)$ with alignments

$$P(f, a|e) = P(a|e)P(f|a, e)$$

$$P(f|e) = \sum_a P(f, a|e)$$

where

$$\begin{aligned} P(f|a, e) &= \prod_j t(f_j|e_{a_j}) \\ P(a|e) &= \prod_j d(a_j|j, l, m) \end{aligned}$$

IBM Model 3

Introduce a fertility ϕ

Example

Mary	did	not	slap	the	green	witch	
1	0	1	3	2	1	1	(chose with $n(\phi_i \mid e_i)$)

Mary	not	slap	slap	slap	the	the	green	witch
Mary	no	daba	una	botefada	a	la	verde	bruja

Model

$$P(f, a|e) = \prod_{i=0}^{l} t(f_{a_j}|e_i) \cdot \prod_{i=1}^{l} n(\phi_i|e_i) \cdot \prod_{a_j \neq 0, j=1}^{m} d(a_j|i, l, m)$$

$$\cdot \prod_{i=0}^{l} \phi_i! \cdot \frac{1}{\phi_0!} \cdot \begin{pmatrix} m - \phi_0 \\ \phi_0 \end{pmatrix} \cdot p_1^{\phi_0} \cdot p_0^{m-2\phi_0}$$

See workbook by Kevin Knight [pdf]

Parallel corpus and non-parallel corpus

Parallel

- Aligned sentence-by-sentence
- Traditional MT: estimate parameters with parallel corpora by using EM

Non-parallel

- No alignment
- Unsupervised learning: decipherment

Word Substitution Decipherment

Properties

- Word-to-word
- Deterministic
- No-reordering

Example

The $\rightarrow$ crqq, saw $\rightarrow$ fxyy, ran $\rightarrow$ qdxx

Generative process

1. Generate an English sentence $e = e_1, \ldots, e_n$ with probability $P(e)$
2. Substitute each word e_i with a cipher token c_i with probability $P(c_i|e_i)$

Bayesian Approach

Smart sample-choice selection
Parallelized Gibbs sampling

Advantages

- Efficient training to scale to large data size
- Efficient inference by using incremental scoring of derivations
- There are no memory bottlenecks
- Prior specification allows us to learn skewed distributions

MT as a Decipherment

Given

Foreign text $f = f_1 \dots f_m$ and a monolingual English corpus

Goal

Translate foreign text into English text $e = e_1 \dots e_l$

Model $P(f|e)$ only with monolingual data
Estimate the model parameters θ in order to maximize the probability of f

$$\begin{aligned} \arg\max_{\theta} \prod_{f} P_\theta(f) &= \arg\max_{\theta} \prod_{f} \sum_{e} P_\theta(f, e) \\ &= \arg\max_{\theta} \prod_{f} \sum_{e} P(e) \cdot P_\theta(f|e) \end{aligned}$$

Bayesian Method

Goal

Train IBM Model 3 parameters t, n, d, p without parallel corpus

Distributions

$$\begin{aligned} f_j|e_i, \theta &\sim \text{Mult}(\theta) \\ \theta|\alpha &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

CRP formulation

$$t_\theta(f_j|e_i) = \frac{\alpha \cdot P_0(f_j|e_i) + C_{history}(e_i, f_j)}{\alpha + C_{history}(e_i)}$$

Result

Word Substitution Decipherment

Method	Decipherment Accuracy	
	Temporal expr.	Transtac
EM	87.8	intractable
Iterative EM	87.8	71.8
Bayesian	**88.6**	**82.5**

MT Decipherment

Method	Decipherment Accuracy	
	Time	OPUS
Parallel (MOSES)	5.6 (85.6)	26.8 (63.6)
Decipherment (EM)	28.7 (48.7)	65.1 (19.3)
Decipherment (Bayesian)	34.0 (30.2)	66.6 (15.1)

Discussion

Q. Why did the Bayesian approach underperform?
How can we improve it?