英文Web 文書中のユーザが知らない単語を予測する語義注釈システム

（旧題：語義注釈システムの単語クリックログからの言語能力情報の抽出）

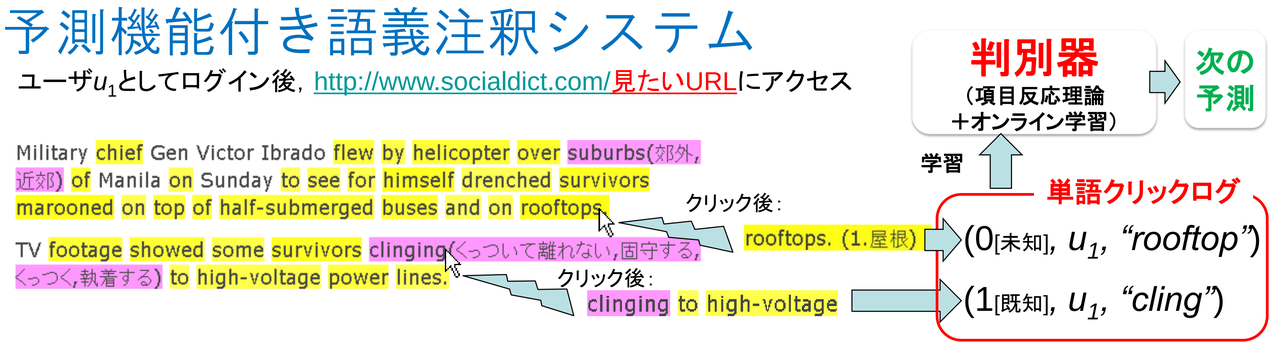
江原　遥，二宮　崇，中川　裕志（東京大学）

要旨

- 英文Web文書中のユーザが知らない単語を予測して，自動的に語義の注釈を行うシステムを作成した
- コーパス別の単語帳を自動作成することも可能
  （ACLやEMNLPの論文の頻出単語のうち自分が知らない語は？）
- TOEFLなどの言語テストで使用されている項目反応理論を使用
- オンライン学習（SGD)で高速化
- 12000語アンケートx16人で評価
- 100個の異なり語程度の学習で精度78.5%．SVMでも約80%でsaturate.

予測機能付き語義注釈システム

ユーザu_1としてログイン後，http://www.socialdict.com/見たいURLにアクセス



赤色：ユーザu_1が知らない[未知]と判定，　黄色：ユーザu_1が知っている[既知]と判定

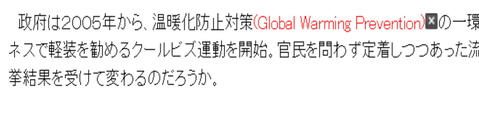
背景

- Web文書の読解支援として，語義注釈システムが提案されている
- 従来の語義注釈システムを拡張し，単語クリックログを解析してユーザが知らない単語を自動的に予測する機能を付与した．

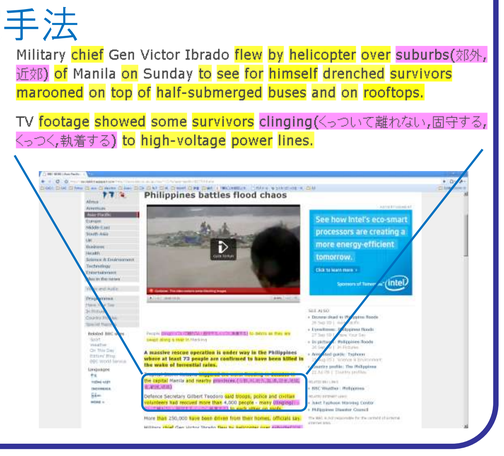
従来手法

pop辞書 (2001)

popIn (2008)

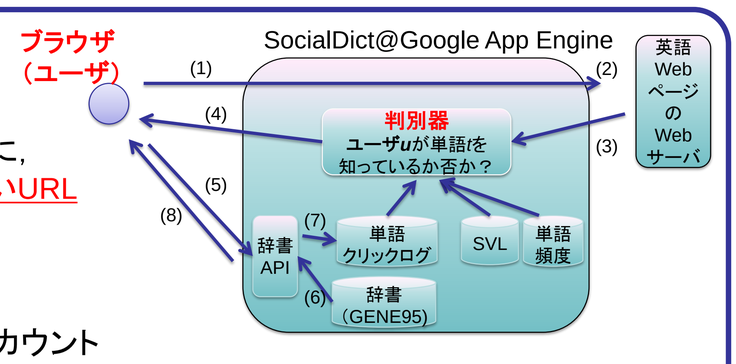


提案手法



システム

- CGI-proxyとして設計
  はてなブックマークなどと同様に，
  http://www.socialdict.com/見たいURL
  でアクセスできるようにした．
- Google App Engineを使用
  高いスケーラビリティ・可用性
  ユーザ登録せずに，Googleアカウントさえあればログインできるようになった



- 付加機能：
  ニュース一覧（BBC)
  単語帳自動作成
  (ACL, EMNLP, etc.)

ログイン

ニュース一覧（BBC)

手法

項目反応理論（item response theory, IRT）：
TOEFLなどの既存の言語テストで利用されているモデルの総称．

Raschモデル：
IRTのうち最も基本的なモデル．
入力（＝蓄積される単語クリックログ）：
$(y_1, u_1, t_1), (y_2, u_2, t_2), \ldots, (y_N, u_N, t_N)$
ユーザ$u \in U$, 単語$t \in T$, $y \in \{0,1\}$
y=1:単語を知っている, y=0:単語を知らない

ロジスティック回帰に帰着

a.k.a.:
対数線形モデル
（Log-linearモデル）
最大エントロピー法

パラメータ：
θ_u:ユーザuの語彙力
d_t :単語tの難しさ

モデル：

$$P(y_n = 1 \mid u_n, t_n) = \sigma(\theta_u - d_t)$$

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$ （シグモイド関数）

とおき，データにフィットするパラメータを最尤推定（又はMAP推定）する．

$$\hat{\boldsymbol{\theta}}, \hat{\mathbf{d}} = \arg\max_{\boldsymbol{\theta},\mathbf{d}} \prod_{n=1}^{N} P(y_n \mid u_n, t_n)$$

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_u, \ldots, \theta_{|U|}), \mathbf{d} = (-d_1, \ldots, -d_t, \ldots, -d_{|T|})$$

$$\mathbf{e}_u = (0, \ldots, \underset{u}{1}, 0 \ldots, 0) \quad (|U|) \qquad \mathbf{e}_t = (0, \ldots, \underset{t}{1}, 0 \ldots, 0) \quad (|T|)$$

$$\mathbf{w}_{rasch} = (\boldsymbol{\theta} \quad \mathbf{d})^{\mathrm{T}}, \boldsymbol{\varphi}_{rasch}(u,t) = (\mathbf{e}_u \quad \mathbf{e}_t)^{\mathrm{T}}$$ とすると，

$$P(y_n = 1 \mid u_n, t_n) = \sigma(\theta_u - d_t)$$
$$= \sigma(\mathbf{w}_{rasch}{}^{\mathrm{T}} \boldsymbol{\varphi}_{rasch}(u,t))$$ と表せる

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \prod_{n=1}^{N} P(y_n \mid u_n, t_n; \mathbf{w})$$

工夫1：素性追加

Rasch:

$$\mathbf{w}_{rasch} = (\boldsymbol{\theta} \quad \mathbf{d})^{\mathrm{T}}$$
$$\boldsymbol{\varphi}_{rasch}(u,t) = (\mathbf{e}_u \quad \mathbf{e}_t)^{\mathrm{T}}$$

LR:
（強化版）

$$\mathbf{w}_{LOGRES} = (\boldsymbol{\theta} \quad \mathbf{d} \quad \mathbf{w}_a)^{\mathrm{T}}$$
$$\boldsymbol{\varphi}_{LOGRES}(u,t) = (\mathbf{e}_u \quad \mathbf{e}_t \quad \boldsymbol{\varphi}_a)^{\mathrm{T}}$$

φ_aに素性を追加することが可能．追加した素性：
- Google 1-gram
  約1兆ページのWebページ中の英単語の頻度．
- SVL12000
  人手で基本的な単語12,000語に対し，難易度を12段階でつけたもの．

工夫2：オンライン学習（Stochastic Gradient Descent, SGD)

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \prod_{n=1}^{N} P(y_n \mid u_n, t_n; \mathbf{w}) = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

$$E(\mathbf{w}) = -\log\left(\prod_{n=1}^{N} P(y_n \mid u_n, t_n; \mathbf{w})\right), -\nabla E(\mathbf{w}) = \sum_{n=1}^{N} -\nabla E_n(\mathbf{w})$$

最急降下法：
（ηは定数）

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta \sum_{n=1}^{N} \nabla E_n(\mathbf{w}^{(k)})$$

SGD:

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta_k \nabla E_n(\mathbf{w}^{(k)}) \qquad \eta_k = \frac{1}{\lambda(k+k_0)}$$ λ, k_0は定数（ハイパーパラメータ）

評価

評価データ作成

どの程度システムを使い込むと，どの程度の精度で予測可能なのかを評価
- 大学生＋大学院生(主に東京大学)16人に，1人12000語について単語を知っている度合いを5段階で自己申告してもらった．
- 辞書引きログがN_0個蓄積されたところにユーザが1人新規にシステムを使い始め，N_1個の単語の既知/未知が得られたと想定．
- この時，そのユーザのTestに含まれる単語のうち何%について既知/未知を正解できたかを1人の精度とした．
- 12000語を表のように分割し、16人の精度の平均値を計算。
- ログは，辞書引きログの代わりに，同じ(y_n, u_n, t_n)の形式のログが残るsmart.fm（旧iKnow)というシステムのログで代用

ログ（N_0)+10～600語(N_1)	Training
1400語	Development
9999語	Test

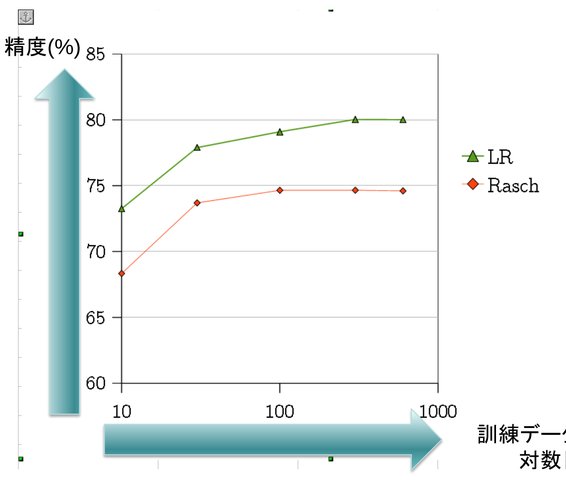
自己申告形式

Dale (1965)	Paribakht+ (1993)	今回の設定	
I never saw it before.	I have never seen this word.	見たこともない	未知
		見たことがある気がする	未知
I've heard of it, but I don't know what it means.	I have seen this word before, but I don't know what it means.	確実に見たことはあるが意味は知らない/覚えたことがあるが意味を忘れている	未知
I recognize it in context.	I have seen this word before, and I *think* it means xxx.	意味を知っている気がする/意味が推測できる	未知
I know it.	I know this word. It means xxx.	意味を確実に知っている	⇨ 既知
	I can use this word in a sentence.		

結果

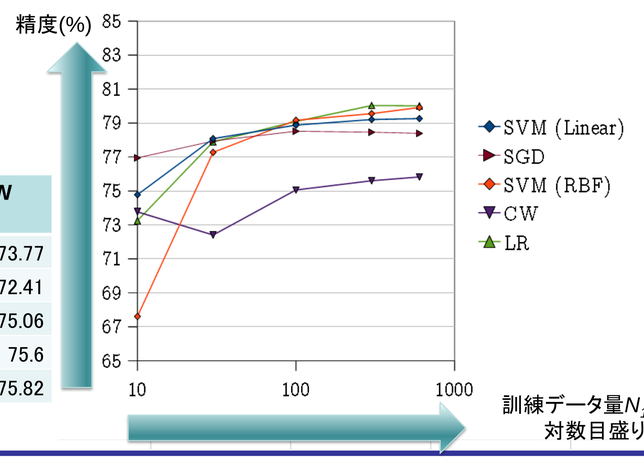
実験1:
素性追加の効果
- 約5%精度向上

	LR	Rasch
10	**73.25**	68.33
30	**77.89**	73.69
100	**79.09**	74.65
300	**80.03**	74.66
600	**80.01**	74.6



実験2:
オンライン学習の効果
- SGD（オンライン学習）はLR（バッチ）に比べて最大約2%の精度減少

	SVM (Linear)	SVM (RBF)	LR	SGD	CW
10	74.78	67.61	73.25	**76.95**	73.77
30	78.08	77.27	77.89	**77.94**	72.41
100	78.88	**79.16**	79.09	78.52	75.06
300	79.2	79.55	**80.03**	78.46	75.6
600	79.27	79.92	**80.01**	78.39	75.82



参考文献

- 豊田秀樹. 2005. 項目反応理論 [理論編]-テストの数理-，朝倉書店.
- Bob Carpenter. 2008. Lazy sparse stochastic gradient descent for regularized multinomial logistic regression. Technical report, Alias-i.
- Yoshimasa Tsuruoka; Jun'ichi Tsujii; Sophia Ananiadou. ACL 2009. *Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty*.