Modèles trigger

Maryse Boisvert
Université de Montréal
Département d'informatique et de recherche opérationnelle

Intelligence artificielle
Traitement statistique des langues naturelles

Plan

1. Introduction
2. Sélection des paires
3. Modèle de langue
   - Interpolation
   - Maximum entropie
4. Résultats

Dépendances locales et à long terme

- N-gramme: prédiction du prochain mot sur la base des prédécesseurs immédiats.
- Cache: la probabilité des M mots les plus récents est augmentée.

$$P_C(w_n|w_{n-1}^{n-M}) = \frac{1}{M}\sum_{m=1}^{M}\delta(w_n, w_{n-m})$$

où $\delta(u,v) = 1$ ssi $u = v$ et $\delta(u,v) = 0$ sinon.

Modèle trigger

- Une paire trigger $a \to b$ est un bigramme à longue distance.
- Rencontrer a dans le texte induit une augmentation de la probabilité de rencontrer b prochainement.
- Capture des dépendances à des distances non-fixées d'avance.

Exemples:

airline ... flights
concerto ... orchestra
asks ... replies

Sélection des paires trigger

- Information mutuelle moyenne entre deux événements pour déterminer l'importance d'une paire:

$$
\begin{aligned}
I(a_0, b) &= P(a_0, b)\log(\frac{P(b|a_0)}{P(b)}) + P(a_0, \bar{b})\log(\frac{P(\bar{b}|a_0)}{P(\bar{b})}) \\
&+ P(\overline{a_0}, b)\log(\frac{P(b|\overline{a_0})}{P(b)}) + P(\overline{a_0}, \bar{b})\log(\frac{P(\bar{b}|\overline{a_0})}{P(\bar{b})})
\end{aligned}
$$

- Calculer l'amélioration en perplexité d'un modèle étendu p_{ab} utilisant la paire $a \rightarrow b$ sur un modèle de base p.

Modèle étendu par backing-off

- Consulter le modèle trigger si la probabilité donnée par le modèle de base n'est pas assez grande.

$$p_{ab}(w|h) = \begin{cases} q(b|a) & \text{si } h \in H_{ab}, w = b \\ (1 - q(b|a)) \cdot \frac{p(w|h)}{\sum_{w' \neq b} p(w'|h)} & \text{si } h \in H_{ab}, w \neq b \\ p(w|h) & \text{si } h \notin H_{ab} \end{cases}$$

où

$$H_{ab} := \{h | a \in h \wedge p(b|h) < p_0\}$$

Modèle étendu par backing-off

- Différence de la log-vraisemblance du modèle trigger et du modèle de base:

$$F(q(b|a)) = \sum_{n:h \in H_{ab}} \log \frac{p_{ab}(w_n|h_n)}{p(w_n|h_n)}$$

- Maximiser cette fonction par rapport à $q(b|a)$.

$$q(b|a) = \frac{\tilde{N}(a;b)}{\tilde{N}(a;b) + \tilde{N}(a;\overline{b})}$$

- Sélectionner les paires dont le modèle étendu démontre la meilleure perplexité.

Modèle multi-trigger

Combiner les K meilleures paires dans un seul modèle.

$$p_T(w_n|h_n) = \frac{1}{|M_n|} \sum_{m \in M_n} \alpha(w_n|w_{n-m})$$

avec:

M_n: ensemble des mots déclencheurs pour la position n

$\alpha(b|a) = \frac{q(b|a)}{\sum_{b'} q(b'|a)}$

(Tillman & Ney 1997)

- Interpolation linéaire entre les modèles trigramme, cache et trigger:

$$p(w_n|h_n) = \lambda_3 p_3(w_n|h_n) + \lambda_C p_C(w_n|h_n) + \lambda_T p_T(w_n|h_n)$$

avec $\lambda_3 + \lambda_C + \lambda_T = 1$.

- Entraînement des paramètres d'interpolation λ par EM.
- Autre méthode de sélection: définir p_{ab} par interpolation linéaire avec un modèle de base.

Maximum entropie

- Estimer la densité de probabilité d'un phénomène aléatoire.
- Information a priori sur cette distribution exprimée sous forme de contraintes.
- Respecter les contraintes tout en laissant le plus de doute possible sur ce que l'on ne connait pas.
- Trouver la distribution $p(y)$ qui maximise l'entropie.

$$H(p) = -\sum_{y} p(y) \log p(y)$$

Maximum entropie

- Définir un ensemble de fonctions caractéristiques $f_i(w, h)$.
- Contrainte: espérance de f_i par rapport à la distribution empirique égale à l'espérance de f_i par rapport au modèle.
- Solution maximum entropie (unique) de la forme:

$$P_{MaxEnt}(w|h) = \frac{\exp^{\sum_i \lambda_i f_i(w,h)}}{Z(h)}$$

Modèle de langue MaxEnt

Fonction caractéristique du trigramme w_1, w_2, w_3

$$f_{w_1,w_2,w_3}(h,w) = \begin{cases} 1 & \text{si } h \text{ se termine par } w_1 w_2 \text{ et } w = w_3 \\ 0 & \text{sinon} \end{cases}$$

Contrainte:

$$\sum_h \tilde{p}(h) \sum_w p(w|h) f_{w_1,w_2,w_3}(h,w) = \frac{1}{N} \sum_{(h,w) \in train} f_{w_1,w_2,w_3}(h,w)$$

Modèle trigger par MaxEnt (Rosenfeld 1996)

$\rightarrow$ Sélection des paires sur la base de l'information mutuelle moyenne.

Fonction caractéristique du trigger $a \rightarrow b$

$$f_{ab}(h,w) = \begin{cases} 1 & \text{si } a \in h, w = b \\ 0 & \text{sinon} \end{cases}$$

• Environ 400000 contraintes trigramme et 35000 contraintes trigger...

Perplexité

Modèles entraînés sur le Wall Street Journal 5 millions de mots et 39 millions de mots. Testés sur 325 000 mots.

modèle de langue	5 M	39 M
trigramme	168.4	104.9
trigramme + cache	138.9	92.1
trigramme + cache + trigger	129.8	88.5
MaxEnt	123	86
MaxEnt + trigramme + cache	108	71

Conclusion

- Amélioration des modèles trigger significative?
- Paires de mots vs paires de classes

Références

- Lau, R. Adaptive Statistical Language Modelling. S.M. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1994.
- C. Tillman, et H. Ney, Word Triggers and the EM Algorithm. Dans T.M. Ellison (ed.) *CoNLL97: Computational Natural Language Learning*, ACL pp 117-124, 1997.
- A. Berger, S. Della Pietra, et M. Della Pietra, A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1), 1996.
- R. Rosenfeld, A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer Speech and Language* 10, 187–228, 1996.
- J. Goodman, A bit of Progress in Language Modeling, *Computer Speech and Language*, pp 403-434, 2001.