Représentation distribuée des mots

Jean-François Paiement
Université de Montréal
Département d'informatique et de recherche opérationnelle
Cours ift6010

27 mars 2003

Résumé

- Indexation par analyse de sémantique latente;
- Décomposition en valeurs singulières (SVD);
- Espace de mots;
- Modèle de langue neuronal probabiliste.

Indexation par analyse de sémantique latente

- On veut situer des documents dans un espace euclidien (Deerwester et al., 1990);
- On construit une matrice des fréquences de mots par document;
- On applique une analyse factorielle bimodale: la décomposition en valeurs singulières (SVD).

Décomposition en valeurs singulières

Toute matrice X de dimension $M \times N$ peut être décomposée en un produit

$$X = TSD^t$$

telles que les colonnes de T et D sont orthonormales et S est diagonale.

On dit que les matrices T et D contiennent les vecteurs singuliers de gauche et de droite et S est appelée la matrice diagonale des valeurs singulières. Cette décomposition est unique à certaines permutations près.

Un algorithme effectuant cette décomposition est décrit dans (Press et al., 1992).

Décomposition en valeurs singulières (suite)

- On conserve les k plus grandes valeurs singulières;
- La matrice $\hat{X}$ est la meilleure approximation de X de rang k (au sens des moindres carrés);
- Choix de la valeur de k;
- Les facteurs d'indexation sont décorellés.

Relations entre deux termes

Le produit scalaire entre deux rangées de $\hat{X}$ reflète le degré d'occurence similaire de deux termes dans l'ensemble des documents du corpus.

Tous les produits scalaires sont contenus dans

$$\hat{X}\hat{X}^t = TS^2T^t$$

On utilise donc TS comme coordonnées des termes dans l'espace. Premier pas vers une représentation plus générale des mots d'une langue naturelle?

Relations document-document et terme-document

- De la même façon,
$$\hat{X}^t \hat{X} = DS^2D^t$$
et DS peut être considéré comme une représentation des documents.
- La comparaison d'un terme et d'un document n'est qu'une entrée de $\hat{X}$.

Traitement de nouveaux documents et de requêtes

- On veut que le traitement d'un document X_i du corpus rende D_i.

  On a donc
  $$D_i = X_i^t T S^{-1}$$
  Correspond à placer le document au centroïde de ses termes;
- Des pré-traitements peuvent être appliqués avant la décomposition afin d'éviter que de plus longs documents aient plus d'influence sur le modèle;
- Une requête est traitée comme une combinaison pondérée de ses termes;
- Problème de polysémie des mots.

Espace de mots (Schütze, 1993)

- Induire une représentation distribuée sémantique pour l'appliquer à la désambiguisation par une méthode de plus proches voisins;
- Pour un vocabulaire de 50,000 mots, il faut conserver 2,500,000,000 comptes de coocurrence;
- On compte plutôt les coocurences de 5,000 4-grammes sélectionnés;
- Il y a 600,000 4-grammes possibles:
  - On enlève tous les 4-grammes rares;
  - Si un groupe de 4-grammes n'apparaît que dans un mot, on n'en conserve qu'un;
  - Les 4-grammes apparaissant dans beaucoup de contextes différents sont enlevés.

Calcul des vecteurs

- Une matrice de collocation des 4-grammes est construite. Chaque entrée $a_{i,j}$ indique le nombre de fois que le 4-gramme i se trouve à moins de 200 4-grammes à gauche du 4-gramme j;
- On extrait k valeurs singulières;
- Pour chaque mot sélectionné, on calcule un vecteur contextuel pour chacune de ses positions dans le texte. Les vecteurs contextuels normalisés et sommés constituent la représentation vectorielle (vecteur de *confusion*) d'un mot.

$$\tau(w) = \sum_{i \in C(w)} \left(\sum_{f \text{ près de } i} \varphi(f) \right)$$

Quelques exemples

word	nearest neighbors
disenchantment	disenchanted sentiment resentment grudging mindful unenthusiastic
kid	dad kidding mom ok buddies Mom Oh Hey hey mama
workforce	jobs employ/-s/-ed/-ing attrition workers clerical labor hourly
disable	deter intercept repel halting surveillance shield maneuvers
Dour	melodies/-dic Jazzie danceable reggae synthesizers Soul funk tunes
Ste.	dry oyster whisky hot filling rolls lean bottle ice
keeping	hoping bring wiping could some would other here rest have

Modèle de langue neuronal probabiliste

On apprend simultanément la représentation distribuée et la fonction de probabilité pour les séquences de mots (Bengio et al., 2003).

Un bon modèle maximise $f(w_t, \ldots, w_{t-n+1}) = \hat{P}(w_t|w_1^{t-1})$ sur l'ensemble de test.

La fonction f est décomposée en deux parties:

1. Une fonction C de n'importe quel élément $i \in V$ vers $C(i) \in \mathbb{R}^m$. Cette fonction est représentée par une matrice de dimension $|V| \times m$;

2. Une fonction $\mathbf{g}_\theta$ telle que $\hat{P}(w_t = i|w_1^{t-1}) = g_{i,\theta}(C(w_{t-1}), \ldots, C(w_t - n + 1))$.

Apprentissage des représentations distribuées

- L'entraînement est réalisé en maximisant la log vraisemblance du corpus d'entraînement:

$$L = \frac{1}{T} \sum_t \log f(w_t, w_{t-1}, \ldots, w_{t-n+1}; \theta) + R(\theta);$$

- Perplexité significativement inférieure aux modèles n-grammes;
- Distribution non-uniforme de la masse de probabilité autour des points d'entraînement.

Conclusion

- Intérêts d'une représentation distribuée des mots:
  - Pour la recherche d'information;
  - Pour la désambiguisation;
  - Pour la représentation d'un modèle de langue.
- Existe-t-il une "vraie" représentation distribuée des mots?

Références

- Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. A Neural Probabilistic Language Model. To appear in *Journal of Machine Learning Research*, 2003.
- S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 (6):391-407, 1990.
- W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, England, second edition, 1992.
- H. Schütze. Word Space. In S. J. Hanson, J.D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages pp. 895-902, San-902, San Mateo CA, 1993. Morgan Kaufmann.