Résumé du Document : Modèles de Langage N-gram

Le document est un chapitre d'introduction aux **modèles de langage N-gram** (LM), une classe fondamentale de modèles d'apprentissage automatique utilisés pour prédire les mots suivants dans une séquence ou pour attribuer une probabilité à une phrase entière.

1. Concepts Fondamentaux des Modèles de Langage

Un modèle de langage est un outil probabiliste qui assigne une probabilité à chaque mot possible suivant une séquence donnée, ou une distribution de probabilité sur l'ensemble des mots possibles.

Utilité des Modèles de Langage :

- **Prédiction de Mots :** La fonction principale est de prédire le mot suivant, ce qui est la base de l'entraînement des grands modèles de langage (LLM).
- **Correction d'Erreurs :** Ils peuvent aider à corriger les erreurs grammaticales ou orthographiques en identifiant les séquences de mots plus probables (par exemple, *« There are »* est plus probable que *« Their are »*).
- **Reconnaissance Vocale :** Ils aident les systèmes à distinguer les séquences acoustiquement similaires mais sémantiquement différentes (par exemple, *« back soonish »* est plus probable que *« bassoon dish »*).
- **Communication Améliorée et Alternative (CAA) :** Ils suggèrent des mots probables pour les utilisateurs qui sélectionnent des mots à partir d'un menu.

2. Le Modèle N-gram

Un **N-gram** est une séquence de N mots. Les termes spécifiques incluent :

- **Unigramme :** Séquence d'un mot.
- **Bigramme (2-gramme) :** Séquence de deux mots.
- **Trigramme (3-gramme) :** Séquence de trois mots.

Le modèle N-gram est basé sur l'**hypothèse de Markov**, qui stipule que la probabilité d'un mot dépend uniquement des $N-1$ mots précédents, et non de l'historique complet de la phrase.

Décomposition de la Probabilité de Séquence : La probabilité d'une séquence de mots $P(w_{1:n})$ est décomposée en un produit de probabilités conditionnelles en utilisant la règle de la chaîne : $P(w1:n) = \prod k = 1nP(wk \mid w1:k-1)$ L'hypothèse de Markov simplifie cette formule pour un N-gramme (où N est la taille du N-gramme) : $P(wn \mid w1:n-1) \approx$

$P(wn \mid wn - N + 1 : n - 1)$ Par exemple, dans un modèle bigramme ($N = 2$), la probabilité d'un mot dépend uniquement du mot précédent : $P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1})$.

3. Estimation des Probabilités

L'estimation des probabilités N-grammes se fait par **Estimation du Maximum de Vraisemblance (EMV)**, qui utilise la fréquence relative des occurrences dans un corpus d'entraînement.

Formule EMV pour un N-gramme : La probabilité d'un N-gramme est calculée en divisant la fréquence observée de la séquence par la fréquence observée de son préfixe : $P(wn \mid wn - N + 1 : n - 1) = C(wn - N + 1 : n - 1)C(wn - N + 1 : n)$Où $C(\cdot)$ est le compte de la séquence dans le corpus.

Considérations Pratiques :

- **Log-Probabilités :** Les probabilités sont stockées et calculées en espace logarithmique (log-probabilités) pour éviter le **sous-débordement numérique** (numerical underflow) qui se produit lorsque l'on multiplie de très petites probabilités. L'addition en espace log est équivalente à la multiplication en espace linéaire.
- **Problème de la Rareté des Données (Data Sparsity) :** Le principal défi des modèles N-grammes est que de nombreuses séquences de mots possibles n'apparaissent jamais dans le corpus d'entraînement, ce qui donne une probabilité de zéro. Cela nécessite des techniques de **lissage (smoothing)** et d'**interpolation** pour réattribuer une partie de la masse de probabilité des événements fréquents aux événements non observés.

4. Évaluation des Modèles de Langage

La performance d'un modèle de langage est évaluée en utilisant un ensemble de test distinct du corpus d'entraînement. La mesure la plus courante est la **Perplexité (PP)**.

Perplexité : La perplexité est une mesure de la qualité avec laquelle un modèle de langage prédit un échantillon de texte. Elle est définie comme l'inverse de la probabilité géométrique moyenne par mot de la séquence de test $W = w_1 w_2 \dots w_N$: $PP(W) = P(w1w2 \dots wN) - N1 = NP(w1w2 \dots wN)1$ Une perplexité plus faible indique un meilleur modèle, car cela signifie que le modèle prédit la séquence de test avec une probabilité moyenne plus élevée. La perplexité est étroitement liée à l'**entropie croisée** du modèle sur les données de test.

5. Conclusion

Le document établit les modèles N-grammes comme une introduction simple et claire aux concepts majeurs de la modélisation du langage, y compris les ensembles d'entraînement et de

test, l'estimation EMV, la perplexité, l'échantillonnage et l'interpolation. Bien que les modèles N-grammes aient été largement remplacés par des modèles neuronaux plus puissants (comme ceux basés sur l'architecture Transformer), ils restent essentiels pour comprendre les principes fondamentaux de la modélisation statistique du langage.