Application des modèles *n*-gramme

- Identification de la langue
  - étant donné un document, identifier dans quelle langue (anglais, français, etc.) il est écrit
- On détermine d'abord un vocabulaire **commun** V pour toutes les langues
- Pour chaque langue l que l'on souhaite détecter
  - on collecte un corpus de documents dans cette langue
  - on assigne une probabilité a priori $P(L=l)$ de la langue
  - on apprend un modèle n-gramme $P(W_i = w \mid w_{i-n+1}, \ldots, w_{i-1}, L=l)$ sur ce corpus
- Étant donné un nouveau document, on lui assigne la langue la plus probable

$$\text{argmax } P(L=l \mid [w_1,\ldots,w_d]) = \text{argmax log } P(L=l,[w_1,\ldots,w_d])$$

$$= \text{argmax log } P(L=l) + \sum_i \log P(W_i = w_i \mid w_{i-n+1}, \ldots, w_{i-1}, L=l)$$

Application des modèles *n*-gramme

- Classification de documents plus puissante
  - l'identification de la langue peut être vue comme de la classification de documents
  - équivaut à remplacer le modèle unigramme du modèle de bayes naïf par un modèle de langage possiblement plus puissant
  - nécessaire si l'ordre des mots est important (« bon » vs. « pas bon »)

- Et plusieurs autres
  - réaccentuation de texte
    - « modele bayesien » → « modèle bayésien »
  - traduction automatique
  - reconnaissance de la parole

Évaluation d'un modèle de langage

- Afin de choisir n, δ ou les λ_i (des hyper-paramètres) on a besoin de définir une notion de performance
  - on choisirait les valeurs qui optimisent cette performance sur un **corpus de validation**, autre que le corpus d'entraînement et de test

- Si on sait dans quel système sera utilisé le modèle de langage, on utilise la performance de ce système
  - ex.: taux de succès d'un système d'identification de la langue

- Sinon, on peut calculer la **perplexité** (perplexité basse= bonne performance)

$$\text{Perp}([w_1,\ldots,w_d]) = \left(P([w_1,\ldots,w_d]) \right)^{-1/d} = \prod_i \left(P(W_i = w_i \mid w_{i-n+1}, \ldots, w_{i-1}) \right)^{-1/d}$$
$$= \exp\left((-1/d) \sum_i \log P(W_i = w_i \mid w_{i-n+1}, \ldots, w_{i-1}) \right)$$

Échantillonner d'un modèle n-gramme

- Pour avoir une idée de la qualité d'un modèle de langage appris, on peut aussi échantillonner de nouveaux documents
  - on laisse la machine parler d'elle-même

- Voici des échantillons de modèles unigramme, bigramme et trigramme, appris à partir du livre de référence

unigramme: « logical are as are confusion a may right tries agent goal the was... »

bigramme: « systems are very similar computational approach would be represented... »

trigramme: « planning and scheduling are integrated the success of naive bayes model is... »