Reconnaissance d'écriture et de la parole

Damien Nouvel

Plan

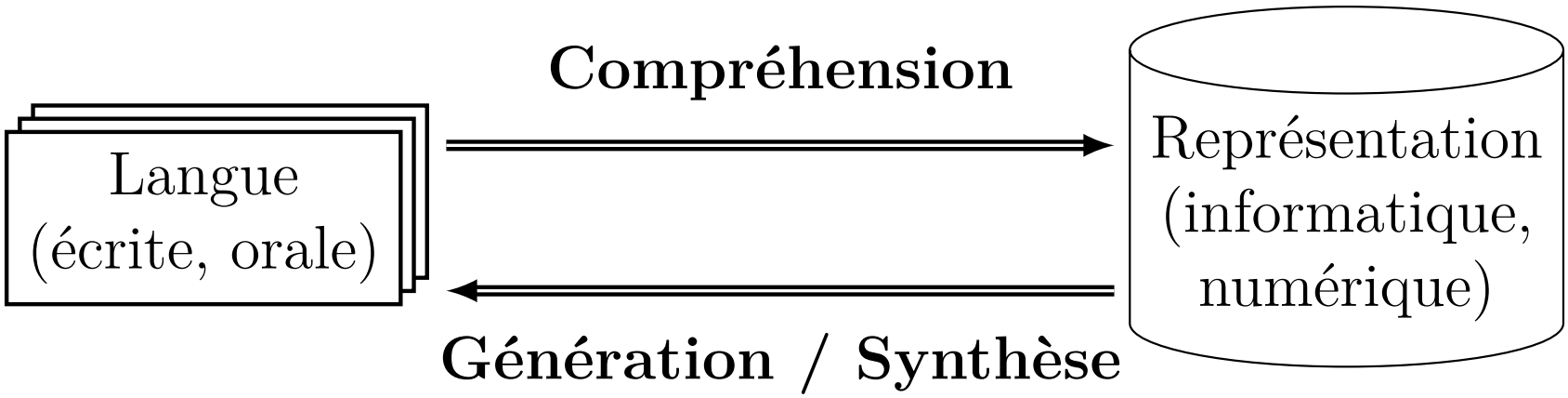
1. Supports, numérisation, reconnaissance

2. Reconnaissance de l'écriture

3. Reconnaissance de la parole

4. Discussions

Génération / compréhension



Langues et supports

- Supports traditionnels (avant l'informatique)
  - Écrit : papier
  - Oral : disques (vynil, CD, cassettes)
- Numérisation du support
  - Écrit : images (JPG, PNG, GIF…)
  - Oral : format audio (MP3, AAC, WMA, MPEG…)

$\Rightarrow$ Objectif : transformer la donnée numérique en langage

$\Rightarrow$ Appel à des méthodes TAL (entre autres)

Processus

- Processus général
  - Acquisition des données
  - Analyse du format en entrée
  - Utilisation de descripteurs adéquats
  - Exploration de l'espace solutions plausibles
  - Décision selon la vraisemblance (modèles de langue)
- Sollicite plusieurs disciplines, en particulier
  - **Mathématiques** (traitement du signal, probabilités)
  - **Informatique** (formats de données, algorithmes)
  - **Linguistique** (plausibilité des reconnaissances)

$\Rightarrow$ Données $\rightarrow$ hypothèses $\rightarrow$ solutions

$\Rightarrow$ Utilisation de méthodes plus ou moins **supervisées**

$\Rightarrow$ Paradigme : **inférence** de modèles à partir de données

Plan

1. Supports, numérisation, reconnaissance

2. Reconnaissance de l'écriture

3. Reconnaissance de la parole

4. Discussions

Acquisition des données

- **Modalités** d'écriture
  - Manuscrite (cursive)
  - Imprimée (fontes / polices)
  - Présence de lettrines / enluminures

  $\Rightarrow$ Complexités selon la langue
- **Types** de documents à traiter
  - Format : courriers, ouvrages, journaux...
  - Thématiques spécifiques (vocabulaire / lexique)
  - Plus ou moins anciens (dégradations, diachronie)
  - Autres éléments para-langagiers (formules, schémas, images...)
- **Encodage** : scan, pages composée de pixels

Traitements

- Traitement du signal
  - Ajustement des **couleurs** (suppression) et du contraste
  - Détection de **blocs** (colonnes, paragraphes, figures, tableaux…)
  - Localisation des **glyphes** (segmentation en caractères)
- Construction d'hypothèses
  - Définition de **descripteurs** (caractéristiques / features)
  - Recensement des **glyphes** possibles pour un alphabet

  ⇒ Écriture (>100 langues écrites, 135 scripts Unicode en 2016)
  - Approche par **similarités**

  ⇒ **Hypothèses** de reconnaissance (graphe)
- Désambiguisation
  - Vérification de la **vraisemblance** par **modèles de langue**

  ⇒ Modèles **markoviens**, **n-grammes**

Applications

- Technologies assez matures
  - Tri postal
  - Traitement de chèques
  - Scans de lettres (courriers)
  - Numérisation de livres (Google Livres)
  - Manuscrits anciens
  - Captcha (apprentissage de modèles)
  - Systèmes de lecture pour aveugles (documents, panneaux)
  - ...

TP OCR

- Exercice
  - Prendre une photo d'un document avec son smartphone
  - S'envoyer la photo par email et l'enregistrer
  - Installer Tesseract sur Ubuntu

    `sudo apt-get install tesseract-ocr`
  - Installer une langue `lng` (parmi plus de 100)

    `sudo apt-get install tesseract-ocr-lng`
  - Exécuter l'OCR pour la langue

    `tesseract doc.tif doc -l lng`

  ⇒ Consulter le résultat du fichier `doc.txt`

  ⇒ Calculer le pourcentage de mots erronés

  ⇒ Discuter les possibles causes d'erreurs

Plan

1. Supports, numérisation, reconnaissance

2. Reconnaissance de l'écriture

3. Reconnaissance de la parole

4. Discussions

Acquisition des données

- Parole humaine (langue, accent…)

⇒ Peu de variation de formats : **audio**

- Documents à traiter
  - Plus ou moins bruités
  - Discours vs conversations
- **Encodage** audio : son compressé
  - Formats variés
  - Avec ou sans pertes

Traitements

- Traitement du signal
  - Acoustique (débruitage)
  - Transformée de Fourier
  - Extraction des **formants**
- Construction d'hypothèses
  - Définition de **descripteurs** (caractéristiques / features)
  - Délimitation des **syllabes** (segmentations possibles)
  - Recensement des **prononciations** possibles

  $\Rightarrow$ Dialectes (5000 langues orales **vivantes**)
  - Approche par **similarités**

  $\Rightarrow$ **Hypothèses** de reconnaissance (graphe / saucisse)
- Désambiguisation
  - Vérification de la **vraisemblance** par **modèles de langue**

  $\Rightarrow$ Modèles **markoviens**, **n-grammes**

Applications

- Technologie aujourd'hui mature (WER acceptable)
  - **Smartphones**
  - Serveurs vocaux
  - Robotique (interaction vocale)
  - Dactylographie
  - Renseignement
  - …

Plan

1. Supports, numérisation, reconnaissance

2. Reconnaissance de l'écriture

3. Reconnaissance de la parole

4. Discussions

Des données et du langage

- Applications très gourmandes en données
  - Processus d'**apprentissage artificiel**
    - Réseaux de neurones
    - Utilisation de l'entropie
    - Itérations d'apprentissage
    - ...

  ⇒ Lien avec l'apprentissage **humain** ?

  ⇒ Intelligence artificielle
- Exploitation de **descripteurs**
  - Processus de **discrétisation** (numérisation)
  - Pas nécessairement **universels** (langues)
  - Nombreux par combinatoire des données

Niveaux d'analyse TAL

- Modules généralement implémentés en TAL
  - **Signal** : audio / image
  - **Morphologie** : phonèmes, syllabes / caractères, morphèmes
  - **Syntaxe** : dépendances entre mots, énoncés
  - **Sémantique** : sens et représentation
  - **Pragmatique** : cohérence discursive

$\Rightarrow$ Interdépendance entre toutes les strates

$\Rightarrow$ Idéalement, travaillent de concert (traitements joints)

$\Rightarrow$ En pratique, séquentiels (pipeline, chaîne de traitement)

TP Voyelles

- Récupérez un texte sur Internet (par ex. article Wikipedia)
- Faire un programme qui enlève les voyelles de ces textes
- Faire un programme pour rétablir les voyelles correctes
  - À l'aide d'un dictionnaire (arbitraire / heuristique)
  - Par méthode probabiliste (mots les plus fréquents)
  - Par méthode contextuelle (bigrames)
  - Par méthode générative (HMM)
- Vous pouvez utiliser les ressources disponible sur `http://redac.univ-tlse2.fr/corpus/wikipedia.html`