Identification et sélection statistiques d'unités lexicales dans les langues peu dotées pour la reconnaissance automatique de la parole

Thomas Pellegrini et Lori Lamel

Groupe Traitement du Langage Parlé
LIMSI-CNRS
http://www.limsi.fr/tlp/

Introduction

Contexte :

- Reconnaissance de la parole pour des langues peu dotées

Principal axe de travail :

- Traiter le problème du manque de textes : taux OOV très élevés, estimation problématique des modèles de langage
- Utilisation de méthodes statistiques pour aider à détecter et sélectionner des unités lexicales adaptées à la tâche de transcription automatique

Cas d'étude :

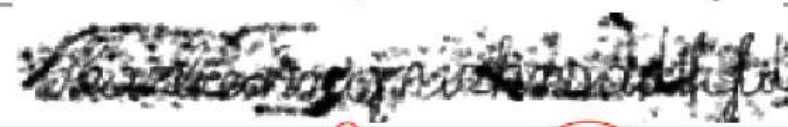
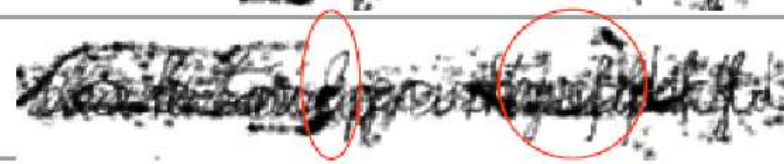
- Amharique, Broadcast News
- Turc, premières études avec parole lue, mais BN visé

Unités lexicales (1/3)

- Notion de mot difficile à définir : mot graphique, mot sonore, mot sémantique, mot lexical ?
- Les mots sont difficiles à identifier :
  - À l'oral, pas de pause entre les mots
  - À l'écrit :
    - Présence éventuelle de séparateurs comme l'espace ou les ponctuations (chinois, japonais pas de séparateur), mot graphique
    - Même avec séparateur de mots l'identification est difficile
  - Exemples :
    - A little bit of, a lot of
    - Kindermädchen (Martinet, éléments de linguistique générale, 1970)
    - Pomme de terre

Extrait cours Tanja Schultz

Why Is Speech Recognition so Difficult?

written text:	Why is speech Recognition so Difficult?
spontaneous:	why's speech recognition so difficult
continuous:	whysspeechrecognitionsodifficult
pronunciation:	whazbeechregnizhnsadifcld
acoustic variability:	
noise:	
Cocktail party-Effect:	

Unités lexicales (2/3)

- Lexiques de reconnaissance formés à partir de
  - Corpus de textes
  - Lexiques disponibles
  - + ajout éventuel de syntagmes très fréquents (co-articulation et réduction inter-mots. Exemple : give_me gimme)
- Langues peu dotées : en général taux d'OOV très élevés, beaucoup de n-grammes peu représentés
- Problème d'OOV similaire avec les langues qui forment les mots par composition de morphèmes
- Question : peut-on définir les unités lexicales de reconnaissance de manière à réduire l'OOV ?
  - Chercher une méthode la plus indépendante possible de la langue, avec le minimum de connaissances linguistiques spécifiques

Unités lexicales (3/3)

- Méthode statistique non-supervisée de décomposition en morphèmes ou morphes, basée sur un corpus
- Exemple : déjouer, rejouer, surjouer ⟶ dé+, jouer, re+, sur+
- Études de décomposition en RAP : allemand (Adda, Eurospeech 2003), arabe (Xiang, ICASSP 2006), turc et finnois (Kurimo, Interspeech 2006), Coréen (Schultz, ICSP 1999)
- Algorithmes de décomposition : Harris (1955), Goldsmith (2001), Morfessor (2005)
- Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0, M.Creutz and K.Lagus, Computer and Information Science, Report A81, 2005
- http ://www.cis.hut.fi/projects/morpho/

Principe général de Morfessor 1.0

Deux modes du programme :

1. un mode “entraînement” de type maximisation a posteriori (MAP) qui utilise des propriétés exprimées sous forme de probabilités ou pseudo-probabilités comme par exemple la probabilité des séquences de caractères
2. un mode “décodage” de type Viterbi pour décomposer un lexique de mots à partir d’un modèle appris. Les mots OOV peuvent être décomposés

1 évalua	+ tion	
1 évalua	+ tion	+ s
1 évalue		
1 évalue	+ nt	
1 évalue	+ r	
1 évalue	+ ra	
1 évalu	+ on	+ s
1 évalu	+ é	
1 évalu	+ ée	
1 évalu	+ ée	+ s

Création d'un modèle de décomposition

- Soit le corpus de textes $\mathbf{corpus}$ (lexique avec comptes éventuels), on cherche le lexique $\hat{L}$ tel que :

$$\underset{\mathsf{L}}{\mathbf{argmax}}\ P(\mathsf{L}|\mathbf{corpus}) = \underset{\mathsf{L}}{\mathbf{argmax}}\ P(\mathbf{corpus}|\mathsf{L})P(\mathsf{L}) \quad (1)$$

- La vraisemblance est :

$$P(corpus|L) = \prod_{i=1}^{W} \prod_{j=1}^{n_i} P(m_{ij}) \quad (2)$$

$$P(m) = f_m/N \quad (3)$$

Entraînement

- Probabilité a priori :

$$P(\mathsf{L}) = P(f_{m_1}, ..., f_{m_M}) P(s_{m_1}, ..., s_{m_M}) \tag{4}$$

$$P(s_{m_1}, ..., s_{m_M}) = P(s'_{m_1}, ..., s'_{m_M}) P(l_{m_1}, ..., l_{m_M}) \tag{5}$$

avec

$$P(s'_{m_i}) = \prod_{j=1}^{l_i} P(c_{ij}) \tag{6}$$

$$P(c_{ij}) = f_{c_{ij}} / N_c \tag{7}$$

et

$$P(l_{m_i}) = (1 - P(</w>))^l P(</w>) \tag{8}$$

Modifications apportées

1. Option de taille minimale de morphes
2. Intégration d'une probabilité de frontière de morphe inspirée de l'algorithme de Harris
3. Intégration de propriétés phonético-acoustiques orientées pour la RAP :
   - Contrainte de confusion entre morphes (alignements préalables)
   - Propriété basée sur des traits distinctifs binaires

Frontière de mot - Harris

- Zellig S. Harris, From Phoneme to Morpheme, Language, volume 31, pp190-222, 1955
- Exemple : Verkehrsabteilung
- Probabilité qui remplace le terme $P(l_{m_i})$ (WB pour “Word Boundary”) :

$$P_H(\mathrm{WB}) = L(\mathrm{WB})/L \tag{9}$$

où $L(\mathrm{WB})$ est le nombre de caractères qui complètent m_i pour former des mots du lexique et L le nombre de caractères distincts de la langue.

- Différence avec Harris : Harris considère la variation de $L(\mathrm{WB})$
- Cette expression favorise a priori les préfixes courts

Contrainte de confusion entre morphes

- Des alignements phonémiques des données audio transcrites ont permis de connaître quelles sont les paires de voyelles qui se substituent le plus l'une à l'autre [Pellegrini et Lamel, LREC 2006]
- Si un morphe ne présente qu'une voyelle de différence avec un morphe du lexique et que les deux voyelles forment une paire de "confusion", alors la décomposition est interdite
- Exemple en amharique : ?xnxdE et ?xnxdi

Propriété phonétique

- Basée sur les traits distinctifs qui sont des propriétés phonologiques abstraites servant à discriminer des phonèmes entre eux
- Propriétés : articulatoires (lieu, voisement, ouverture du conduit vocal, etc. . .) et acoustico-perceptives (trait grave/aigu)
- R.Jakobson (Preliminaries to Speech Analysis, 1952) : les phonèmes d'une langue sont différentiables par un nombre limité de traits, ces traits pouvant être binaires ou univalents
- Un système de traits pour décrire une langue n'est pas unique ni forcément bien défini
- But : favoriser les morphes qui ont des traits distinctifs les plus différents possible

Exemple traits distinctifs des voyelles turques

Trait	voyelles							
Phone LIMSI	a	e	i	u	y	x	o	@
Graphème	a	e	i	u	ü	ı	o	ö
symbole API	a	e	i	u	y	ɪ	o	ø
haute	0	0	1	1	1	0	0	0
basse	1	1	0	0	0	0	1	1
arrondie	0	0	0	1	1	0	1	1
réduite	0	0	0	0	0	1	0	0
antérieure	0	1	1	0	1	0	0	0
longue	1	1	1	1	1	0	1	1

Intégration à l'algorithme Morfessor (1/2)

- Ajout d'un terme à l'équation générale

$$P(\mathrm{L}) = P(f_{m_1}, ..., f_{m_M}) P(s_{m_1}, ..., s_{m_M}) D(td_{m_1}, ..., td_{m_M}) \quad (10)$$

- Avec

$$D(td_{m_1}, ..., td_{m_M}) = \prod_{k=1}^{M} D(m_k) \quad (11)$$

- Calcul de $D(m_k)$ restreint à la comparaison des traits distinctifs du morphe m_k avec ceux des morphes qui ont même racine consonantale pour les traits des voyelles et même suite de voyelles pour les traits des consonnes

$$D_{td}(m_k) = \prod_{j=1}^{j=N_k-1} D_{td}(m_k, m_j) \quad (12)$$

Intégration à l'algorithme Morfessor (2/2)

$$D_{td}(m_k, m_j) = \prod_{l=1}^{l=V_k} \frac{\Delta_{kl,jl}}{C} \quad (13)$$

- N_k : nombre de morphes qui partagent la même racine consonantique
- $\Delta_{kl,jl}$: nombre de traits différents de la l^{me} voyelle des morphes m_k et m_j
- C : nombre total de traits différents considérés

Application à l'amharique (1/3)

Source	Train	Dev
Deutsche Welle	24h06	1h20
Radio Medhin	11h08	0h37
# Locuteurs	200	15
# Mots	232.6k	14.1k
# Mots distincts	45.8k	5.8k

- Textes récupérés sur Internet : 4.6M mots, 112.7k mots distincts

Application à l'amharique (2/3)

- Système de référence : deux passes avec une adaptation non-supervisée des modèles acoustiques après la première passe (J-L.Gauvain, The LIMSI Broadcast News Transcription System, 2002)
- 10.5k contextes différents, modèles dépendants de la position dans les mots (3 états par modèle), avec un total de 8.5k états liés (32 gaussiennes par état). ML 4g
- Jeux de modèles spécifiques à chaque modèle de découpage :

Option	*Signification*
BL	Système baseline basé sur les mots, pas de décomposition
M	Morfessor 1.0 baseline
M H	M + modification 'Harris'
M H DF	M H + paramètre traits distinctifs (voyelles)
Cc	+ contrainte de confusion

Performances (3/3)

Options	*# Morphes*	*OOV (%)*	*WER (%)*
BL	133384	6.9	24.0
M	95937	4.3	24.1
M Cc	128239	4.6	23.9
M H	90740	4.2	24.5
M H Cc	126105	4.5	23.7
M H DF	94198	4.2	24.3
M H DF Cc	128404	4.5	23.6

- Système de référence : OOV=6.9% soit 971 mots
- Système M H DF Cc : 69/971 mots sont OOV, 902 mots décomposés.
- 123/902 mots (14%) correctement reconnus soit 1% de gain sur les 14k mots de test
- Gain observé de 0.4%

Premières expériences en turc (1/2)

Source	Train	Dev
LDC	5h	0h30
# Locuteurs	108	12
# Mots	30.5k	3.3k
# Mots distincts	6.7k	2.1k

- Textes récupérés sur Internet : 173k mots, 27.5k mots distincts
- Grande disparité de corpus entre l'audio (phrases phonétiquement équilibrées) et les textes de news provenant d'Internet

Premières expériences en turc (2/2)

Options	*# Morphes*	*OOV (%)*	*WER (%)*
BL	32k	28.3	50.5
M	15.7k	23.9	58.6
M H	13.7k	23.0	59.4
M H Cc	19.3k	24.5	57.0

- Grapheme error rate : BL 20.5% > M H Cc 19.4%
- Problème dans la reconstitution des mots

REF :güneş	*	hareketleri	herhalde uzun süren	seller	ya da	kuraklık getirir
HYP :güneş	hareket	leri	herhalde uzun süren	seler	ya da	* kuraklıkgetiri
Eval :	I	S		S		D S

Conclusion, perspectives

- Paradigme statistique de décomposition des mots avec de nouvelles propriétés destinées à la tâche de reconnaissance
- Amharique : léger gain observé sur un corpus BN pour l'amharique où des mots qui étaient OOV ont été correctement reconnus
- Turc : légère dégradation des performances sur un très petit corpus de parole lue en turc. Très fort taux d'OOV. Résultats très préliminaires certainement peu fiables pour l'instant
- Question de la recomposition des mots : pour éviter d'avoir des mots qui sont également des préfixes, utiliser une balise de fin de mot plutôt qu'un caractère acollé au préfixe
- Collecte en cours de textes en turc pour diminuer l'OOV

Merci !