Extraction d'entités nommées à partir de graphes de mots

Frédéric Béchet

Aix Marseille Université
LIF-CNRS Laboratoire d'Informatique Fondamentale de Marseille

Introduction

- **Pourquoi utiliser des graphes de mots ?**
  - Représentation d'une entrée ambiguë
    - Texte « non natif », généré par un processus automatique
      - Reconnaissance Automatique de la Parole (RAP)
      - Reconnaissance de caractères
      - Traduction automatique
      - Résumé automatique
    - Ambiguïtés de segmentation dans des textes bruts (et bruités)
      - Segmentation en phrases
      - Segmentation en tokens
- **Qu'est ce que ça change ?**
  - Recherche jointe
    - Meilleure séquence de mots ou « tokens »
    - Meilleure séquence d'entités nommées
  - Problème
    - Limiter l'explosion combinatoire

Introduction

- **Deux approches complémentaires**
  - Evaluation jointe des chemins tokens / entités nommées
    - Modification de la fonction de coût d'un chemin
    - Recherche du meilleur chemin tokens/entités
  - Recherche d'entités dans le graphe d'hypothèses
    - Transformation du graphe d'hypothèses
      - Probabilités a posteriori, mesures de confiances
      - Réseaux de confusion
    - Utilisation d'informations a priori
      - Définir les entités potentielles et leurs structures
    - Vérification de la présence des entités dans le graphe
      - score de confiance pour chaque détection

Introduction

- **Cadre de ces travaux**
  - Reconnaissance d'entités nommées dans des flux audio
    - Traitement des disfluences, des erreurs de reconnaissance
  - Corpus
    - Dialogues Homme-Machine
      - Corpus How May I Help You ? (AT&T)
    - Emissions radiophonique
      - Corpus ESTER
    - Dialogues enregistrés dans des centres d'appels
      - Corpus DECODA

Reconnaissance d'EN dans des flux audio

- **Problématique**
  - Parole spontanée = disfluences
    - Exemple : corpus DECODA
      - Centre d'appel de la RATP

- **rue euh Bretagne**
- **en fait on on prend euh on part par euh l' aéroport euh Roissy - Charles de Gaulle**
- **euh là je suis à l' arrêt euh Trosy**
- **et après vous avez la** rue Cavai Cavillon
- **il est au bout de la rue euh Jean c' est avenue hein Jean Mermoz**
- **gare euh Montparnasse**
- **de la gare des des Ardoines**
- **l' arrêt Louis euh Boilly**
- **en vérité il est rue du Colonel a Avia c' est le long de euh**
- **Porte de d' Orléans**
- **et du six décembre au au six au vingt-six décembre**
- **c' est pendant cinq semaines depuis le deux deux novembre**
- **le premier premier février**
- **mardi dix euh novembre**

DECODA

Reconnaissance d'EN dans des flux audio

- **Texte non natif**
  - Génération d'une transcription W à partir d'un signal A :

$$P(\hat{W}|A) = \max_{W} P(W|A) \qquad P(W|A) = \frac{P(A|W)P(W)}{P(A)}$$

  - Principale conséquence ⇨ Vocabulaire fermé
    - Tout les mots doivent être dans le modèle génératif P(W)
    - Les transcriptions automatiques ne contiennent aucun mot inconnus !!
  - Segmentation en phrases
    - Silences, fixe, indices prosodiques/lexicaux/syntaxiques
  - Disfluences
    - Très mal transcrites
  - Hypothèses multiples avec des scores de confiance
    - Graphes de mots, réseaux de confusion, ….

Extraire des entités nommées d'archives sonores

- **RAP donne des résultats corrects à condition ..**
  - Très grande similitude entre les corpus d'apprentissage et de test
- **Cependant ..**
  - Archives sonores de grande taille
    - Très grande variété de domaine, grande période temporelle
    - Décalage Thematique / Temporel
- **Performance RAP / EN**
  - WER and F-measures sont corrélées (*Miller et al., ANLP-NAACL 2000*)
  - Les performance NE sont très affectées par la présence de mots hors-vocabulaire

Système développés pour le texte *vs.* Sorties RAP (1/2)

- **Flux de mots**
  - ni ponctuation, ni découpage en phrase, ni capitalisation
- **Erreurs de RAP**
  - Disfluences, bruit, confusions lexicales
    - ⇒ Modéliser les erreurs de transcription *(Palmer, HLT 2001)*
    - ⇒ Utiliser des treillis de mots *(Saraclar, HLT-NAACL 2004)*
  - Mots hors-vocabulaire
    - RAP optimisée pour les mots les plus fréquents
    - ⇒ Utiliser des méta-données pour mettre à jour les modèles de RAP et d'entités nommées

Système développés pour le texte *vs.* Sorties RAP (2/2)

- **Sur le texte**
  - Point clé => capacité de généralisation pour les mots hors-vocabulaire
    - Contexte, indication graphique, morphologie
    - Avantage: augmenter le rappel Risque: diminuer la précision
- **Sur les sorties RAP**
  - Principale caracteristique ⇨ vocabulaire fermé
    - Les mots inconnus peuvent être détecté et des fois réparé
      - *(Bisani, Eurospeech 2005)*
    - Mais pas de processus générique !!

Pas besoin des capacités de généralisation des systèmes développés sur le texte, utilisation d'informations a priori sur tous les mots du lexique de RAP

Un système développé pour traiter les flux audio

- **Système LIA_NE (participation à ESTER2)**
  - Système téléchargeable (avec modèles) sur ma page WEB du LIF
    - www.lif.univ-mrs.fr
  - Système à 4 niveaux

1. Étiqueteur lexical
   - morpho-syntaxique/sémantique
   - Modèle à base de Hidden Markov Models (HMM)
1. Segmenteur en Entités Nommées (EN)
   - Trouver les frontières et le type des EN
   - Modèle à base de Conditional Random Fields (CRF)
1. Regroupement d'entités, correction de frontières
   - Entités englobées / englobantes
   - Règles manuelles / obtenues sur corpus
1. Validation par base de connaissance
   - Dictionnaire d'entités
   - Classifieurs pour la validation contextuelle

Un système développé pour traiter les flux audio

- **Adaptation au traitement de transcriptions automatiques**
  - Apprentissage sur sortie “bruitée”
    - Suppression des ponctuations et majuscules
  - Collectes de données sur tous les noms propres du lexique de RAP
    - Wikipedia
    - Corpus journalistiques : AFP, Le Monde
    - Transcriptions de l’oral : ESTER, EPAC

- **Traitement des sorties multiples**
  - Pour ESTER 2, pas d’intégration avec le traitement de graphes de mots
    - Traitement des n-meilleures hypothèses uniquement
  - Problèmes d’ambiguïtés
    - N-meilleures séquences de mots
      - Module de Reconnaissance Automatique de la Parole
    - N-meilleures séquences de Parties de Discours (POS)
      - Étiqueteur HMM
    - N-meilleures séquences d’étiquettes Entités Nommées
      - Etiqueteur CRF

Corpus d'apprentissage « brut » et « adapté »

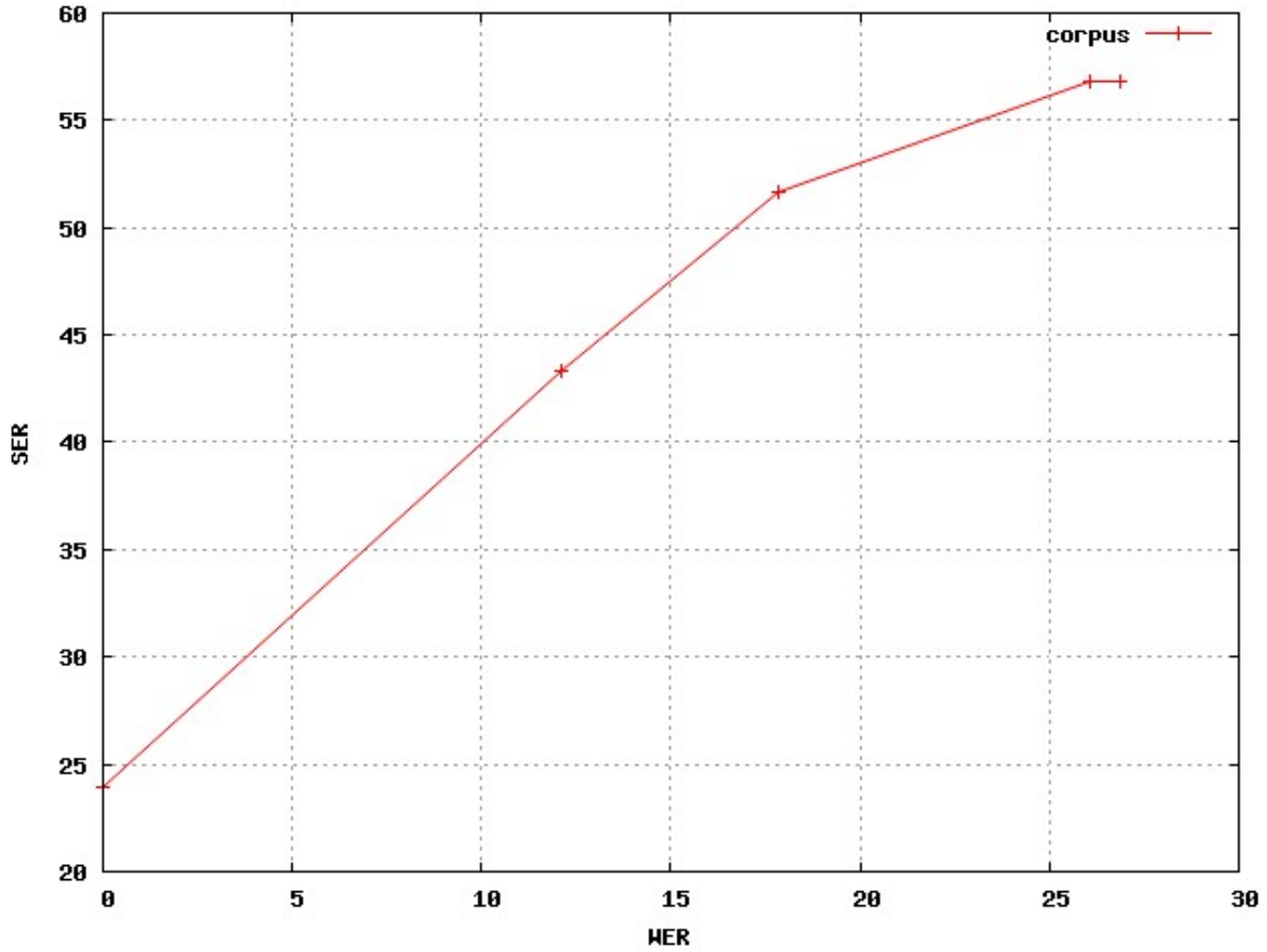
Corpus avec casse et ponctuations

bonjour	NMS	O
.	YPFOR	O
investiture	NFS	O
aujourd'hui	ADV	B-TIME
à	PREPADE	O
Bamako	XLOC	B-LOC
,	YPFAI	O
Mali	XLOC	B-LOC
,	YPFAI	O
du	PREPDU	O
président	NMS	B-FONC
Amadou	XPERS	B-PERS
Toumani	XPERS	I-PERS
Touré	XPERS	I-PERS
,	YPFAI	O
réélu	VPPMS	O
en	PREP	B-TIME
avril	NMS	I-TIME
dernier	AMS	I-TIME

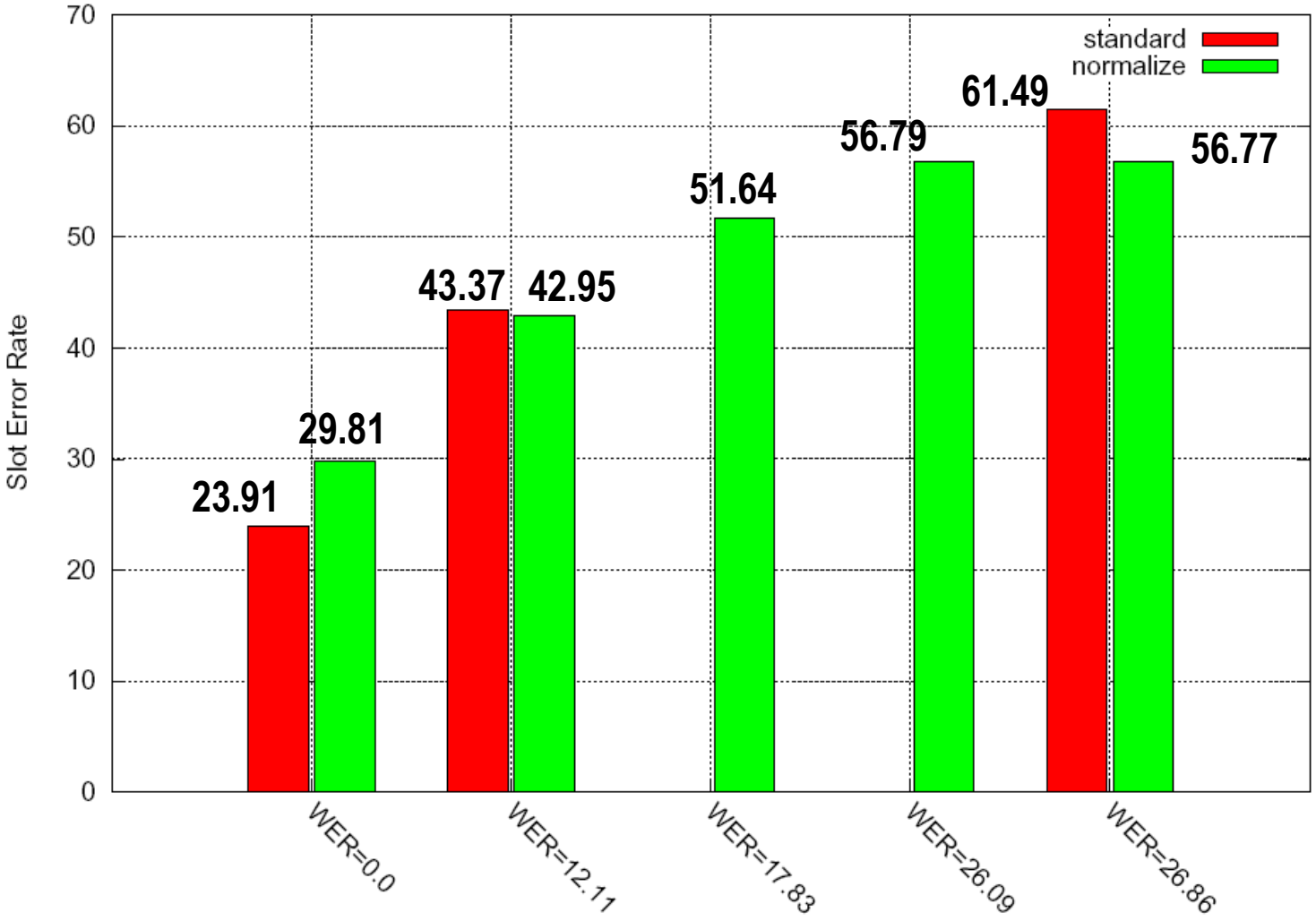
Corpus « normalisé » (pas de ponctuation, tout en minuscule)

bonjour	NMS	O
investiture	NFS	O
aujourd'hui	ADV	B-TIME
à	PREPADE	O
bamako	XLOC	B-LOC
mali	XLOC	B-LOC
du	PREPDU	O
président	NMS	B-FONC
amadou	XPERS	B-PERS
toumani	XPERS	I-PERS
touré	XPERS	I-PERS
réélu	VPPMS	O
en	PREP	B-TIME
avril	NMS	I-TIME
dernier	AMS	I-TIME

Lien Taux Erreur Mots / Erreur Entités Nommées



Influence du formatage des données



Traitement de graphes de mots

- **Intégration des processus**
  - Recherche de meilleurs chemins / reconnaissance EN
  - Modification de la fonction de coût utilisée en RAP

$$\tau(\hat{w_{1,n}}) = \underset{w_{1,n},t_{1,n}}{argmax} P(t_{1,n}, w_{1,n}|A) \approx \underset{w_{1,n},t_{1,n}}{argmax} P(A|w_{1,n}) \times P(t_{1,n}, w_{1,n})$$

    - Le terme $P(t_{1,n}, W_{1,n})$ = modèle de langage + modèle d'étiquetage

  - Implémentation
    - Facile en utilisant des HMM
      - Favre, 2005

$$P_{t_{1,n},w_{1,n}} \approx \prod_{i=1}^{n} P(w_i, t_i|w_{i-1}, t_{i-1}, w_{i-2}, t_{i-2})$$

      - Horlock, 2003

$$P_{t_{1,n},w_{1,n}} \approx \prod_{i=1}^{n} P(w_i|w_{i-1}, w_{i-2}, t_i) \times P(t_i|t_{i-1})$$

Traitement de graphes de mots

- **Problèmes**
  - Modèles de langage mots appris sur de très grandes quantités de données
  - Ressources limitées étiquetées en Entités Nommées
  - Difficulté d'intégration de méthodes discriminantes
    - MaxEnt, CRF, SVM, …
  - Performance limitée de cette approche mais ..
    - Passage d'un graphe de mots à un graphe d'entités nommées
- **Méthodes alternatives**
  - Effectuer d'abord une recherche de meilleurs chemins en mots
    - Graphe avec probabilité a posteriori / Réseaux de confusion
  - Rechercher ensuite les entités dans des zones du graphes
    - Sélectionnées à l'aide d'informations a priori
    - Détectées grâce à des indices dans le graphe de mots
  - Expériences
    - How May I Help You ?
    - ESTER
    - DECODA

How May I Help You ? (AT&T)

- **Méthode**
  - Extraction de connaissances a priori sur le dialogue en cours
  - Exemples:

system>	in Marseille I propose the Hotel la Fanette and the Hotel du Port
user>	where is the Hotel la Fanette?
ASR>	***where is the Hotel Lafayette***

I wanna know why I was charged on **September sixth 11 dollars 63 cents** for calling **8 5 6 2 1 6 5 5 2 1 Clementon New Jersey** for **1 minute**

PHONE BILL SEPTEMBER 2001

DATE	PHONE#	DURATION	PLACE	AMOUNT
09062001	8562165521	01:00	Clementon, NJ	11.63
….	….	….	….	….
….	….	….	….	….

Exemple: AT&T How May I Help You? [tm]

How May I Help You ? (AT&T)

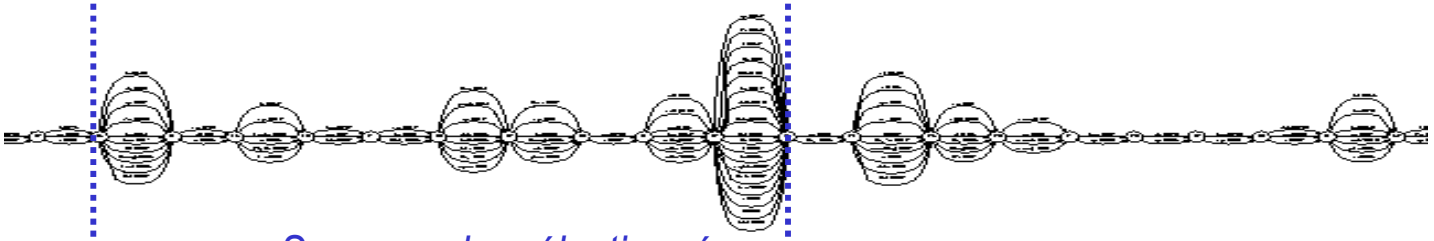
- **Méthode**
  - Détection d'hypothèses d'entités nommées typées sur la meilleure transcription en mot
    - Tagger HMM appris sur des transcriptions automatiques
      - Sur-détection d'entités
  - Recherche des entités dans un réseau de confusion
    - Dans les zones sélectionnées par le tagger
    - En utilisant le type détecté
    - Production d'une liste de n-meilleures hypothèses sur les entités nommées
  - Exemple
  - Référence
    - Dilek Hakkani-Tur and Frederic Bechet and Giuseppe Riccardi and Gokhan Tur Beyond ASR 1-Best: Using Word Confusion Networks for Spoken Language Understanding Computer Speech and Language, Elsevier pages 495-514 volume 20, Issue 4 October 2006 (2006)

How May I Help You ? (AT&T)

Transcr.→ bos I'm Claire Ferguson at **8 5 0 6 3 8 1 8 2 6** I would like to know …

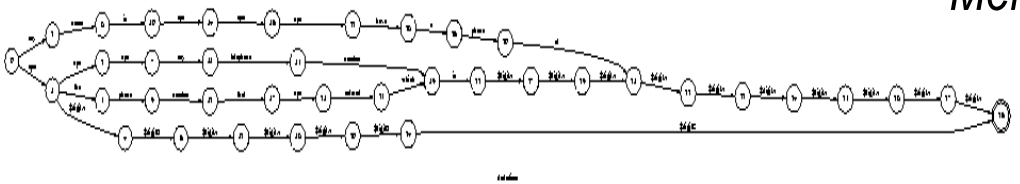
1-best. → bos **<PHONE>** **I'm** *ca-* **8 5 0 6 3 8 1 8** *cents* **</PHONE>** I would like to know

Graphe de mots →



Sous-graphe sélectionné

Composition avec les automates



Meilleurs chemins

N-BEST sur les valeurs

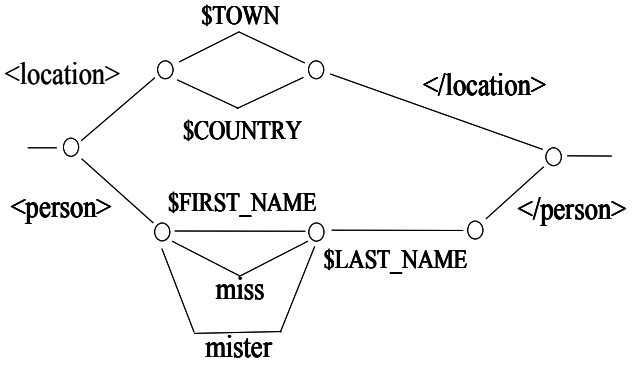
1] **8 5 0 6 3 8 1 8 2 6**
2] 1 8 5 0 6 3 8 1 8 2
3] 8 5 0 6 3 8 1 8 8 6
4] 8 5 0 2 3 8 1 8 2 5

ESTER I

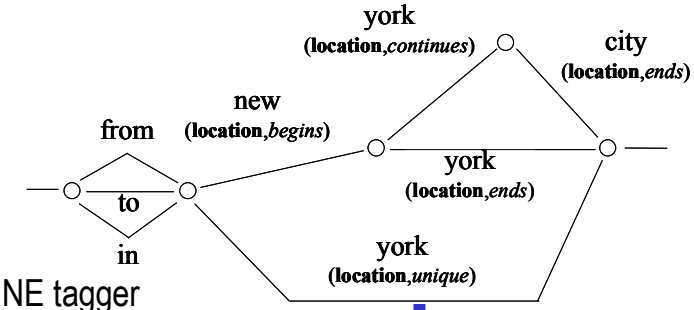
- **Principe**
  - Sélection d'une liste d'entités par rapport à une période temporelle
  - Recherche systèmatique de chaque entité dans le graphe de mots issu du module de reconnaissance de la parole
- **Méthode**
  - Système à 2 niveaux
    - Constitution d'une grammaire des entités nommées
      - Faible longueur des entités
      - Patrons d'entités obtenus sur des corpus étiquetés
      - Grammaires lexicalisées par les mots du lexique de RAP
    - Composition entre le graphe de mots et la grammaire représentée sous forme d'automates
      - Transduction mots ⇨entités
    - Réévaluation des chemins par un étiqueteur HMM
    - Projection du transducteur vers les entités nommées
      - Énumération des n-meilleures hypothèses d'entités nommées
      - Mesures de confiance : proba du système de RAP + scores tagger

ESTER I

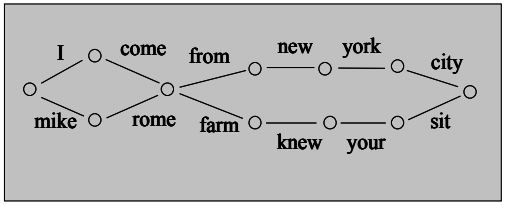
Speech Signal



NE grammars

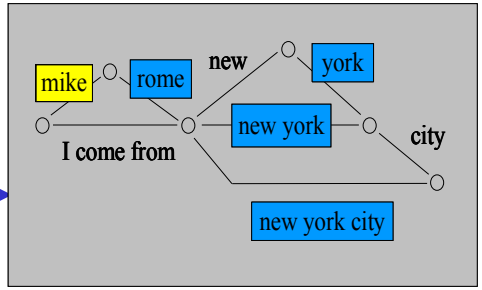


NE tagger



Word Lattice

Composition



Entity Lattice

N-best paths

Composition

Locations

new york city	s_asr
new york	s_asr
york	s_asr
rome	s_asr

Persons:

mike	s_asr

s_tag
s_tag
s_tag
s_tag
-
s_tag

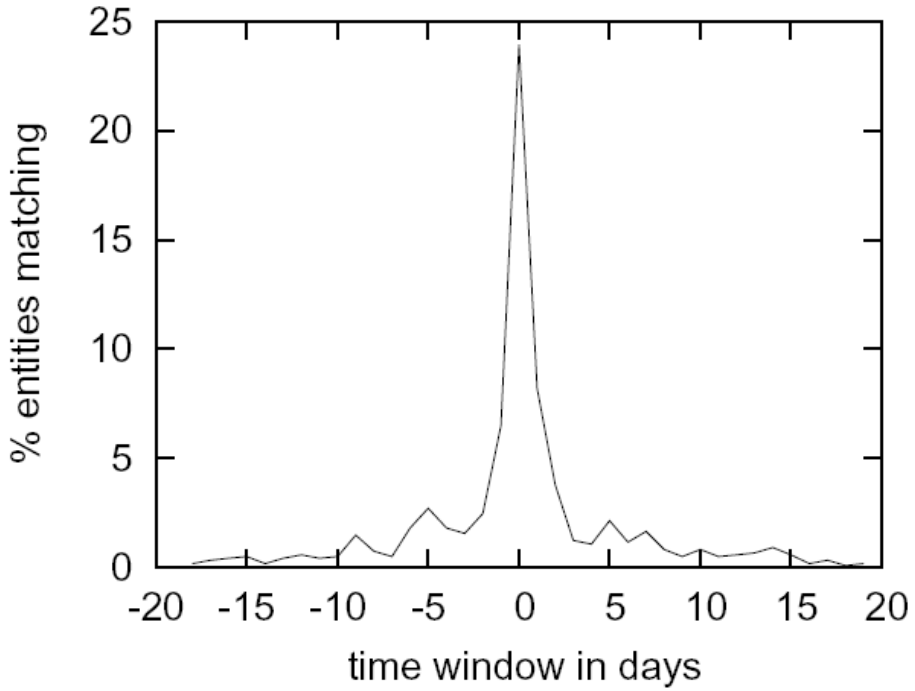
ESTER 1 : exemple de collecte de méta-données

- **Utilisation de l'information temporelle**
  - Meta-données collectées : Newsletters du journal Le Monde ($corpus_{NL}$)
    - Données journalières, du 1er janvier au 31 décembre 2004
    - Couvre la période temporelle de $Test_2$
    - Mais ce ne sont que des résumés d'actualité ..
  - Description du $corpus_{NL}$
    - 1 M de mots avec 140K EN
    - 72% des EN n'apparaisse qu'une fois !! (similaire à *Whittaker, ITRW 2001*)

ESTER 1 : exemple de collecte de méta-données

- **% de EN du $Test_2$ qui apparaissent au moins *n* jours avant ou après dans $corpus_{NL}$**
- **Pic pour *n*=0 mais ..**
  - seulement 25% d'EN communes
  - cependant ..

Ce sont les EN clés pour Caractériser l'actualité d'une journée



ESTER 1 : Evaluation

- **Lexique**
  - 1K nouveaux noms propres ajoutés
    - Réduction du taux de mots hors vocabulaire (OOV) = 0.14% sur $Test_2$
- **Performances de RAP et EN**
  - Petites améliorations
    - WER: 26.4 → 26.1
    - F-measure: 63 → 63.3
    - Oracle recall: 76.9 → 79.9
- **Focus**
  - entités de $corpus_{NL}$ apparaissant le même jour dans le corpus de test

ESTER 1 : Evaluation

- **Résultats**

Condition	Precision	Recall	F-measure	Oracle recall
no adaptation	87.0	75.7	80.9	83.6
avec adaptation	87.5	83.9	85.7	92

  - Amélioration importante de la mesure de rappel
  - Adapter les modèles avec des méta-données
    - Peu d'impact sur les performances globales
    - Mais impact important sur les entités correspondant aux données d'adaptation
  - Référence
    - Benoît Favre and Frédéric Béchet and Pascal Nocéra Robust Named Entity Extraction from Spoken Archives EMNLP (2005) Vancouver, Canada

DECODA

- **Corpus de conversations orales**
  - Projet ANR CONTINT 2009
  - Centre d'appels de la RATP
  - Beaucoup de disfluences mais entités restreintes
    - Lieux, adresses, quantités, produits
  - Annotations en disfluences et entités nommées
  - Principe de traitement
    - Chaîne d'outils MACAON
    - Prendre de l'ambiguïté en entrée
    - Garder l'ambiguïté à chaque niveau de traitement
      - Graphes de mots
      - Graphes de Parties de Discours (POS)
      - Graphes de chunks
      - Graphes d'entités nommées

DECODA

- **Gestion de l'ambiguïté avec MACAON**
  - Chaque niveau génère son propre graphe d'hypothèses
  - Pas de mélange des scores (jusqu'à la décision finale)
  - Chaîne de traitement
    - Graphes de mots
    - Enrichissement du graphe avec la détection des disfluences
      - Rajouts de chemins pour corriger les disfluences « simples »
    - Détection des entités nommées sur le graphe de mots
      - Tagger HMM + tagger CRF
      - Enrichissement du graphe avec les entités potentielles détectées
    - Analyse syntaxique
      - Étiquetage en partie de discours
      - Découpage en chunks
      - Analyse en dépendance

DECODA

- **Gestion de l'ambiguïté avec MACAON**

<table>
<tr><td rowspan="3">syntaxique</td><td></td><td>SN</td><td colspan="2">SN</td><td colspan="2">SV</td></tr>
<tr><td colspan="4"></td><td colspan="2">SP</td></tr>
<tr><td>SN</td><td>SV</td><td colspan="4">SN</td></tr>
<tr><td rowspan="3">morpho-syntaxique</td><td rowspan="3">np</td><td>nc</td><td rowspan="3">det</td><td colspan="3">nc</td></tr>
<tr><td rowspan="2">v</td><td rowspan="2">nc</td><td colspan="2">v</td></tr>
<tr><td>prep</td><td>nc</td></tr>
<tr><td rowspan="3">lexical</td><td rowspan="3">Jean</td><td>lange</td><td rowspan="3">une</td><td colspan="3">pomme de terre</td></tr>
<tr><td rowspan="2">mange</td><td>tome</td><td colspan="2">déterre</td></tr>
<tr><td>pomme</td><td>de</td><td>terre</td></tr>
<tr><td rowspan="2">pré-lexical</td><td rowspan="2">Jean</td><td>lange</td><td rowspan="2">une</td><td>tome</td><td colspan="2">déterre</td></tr>
<tr><td>mange</td><td>pomme</td><td>de</td><td>terre</td></tr>
</table>

Conclusion

- **Traiter des graphes de mots**
  - Intégration des processus de reconnaissance / analyse de parole
  - Gestion des ambiguïtés
    - Disfluences
    - Erreurs de reconnaissance
  - Intégration de connaissances a priori dans le processus de détection
- **Applications**
  - Intégration dans la suite d'outils de TAL MACAON
    - http://macaon.lif.univ-mrs.fr
  - Projet ANR DECODA
    - « speech analytics »
  - Campagne d'évaluation ANR/DGA REPERE
    - Détection de personnes dans des vidéos
  - Campagne d'évaluation ANR ETAPE
    - Emissions télévisuelles
- **Outils**
  - Téléchargement : http://pageperso.lif.univ-mrs.fr/~frederic.bechet