

Extraction d'informations

Reprise de slides de T. Poibeau (CNRS et U. Paris 13), de W. Cohen (CMU), de Julien Lemoine (Exalead)

Antoine Rozenknop

6 février 2009

Sources des Slides

- Thierry Poibeau : (CNRS UMR 7030 et Université Paris13)
Disponible sur <http://www-lipn.univ-paris13.fr/poibeau/laics/>
- W.Cohen (Carnegie Mellon U.)
Disponible sur <http://www.cs.cmu.edu/wcohen/>

1 Les conférences MUC

2 Extraction d'informations avec des transducteurs à états finis

1 Les conférences MUC

2 Extraction d'informations avec des transducteurs à états finis

Première campagne d'évaluation dans le TLN

- *Message Understanding Conferences*: 7 conférences de 1987 à 1998
- Organisées par le DARPA et le NOSC (Naval Ocean System Center)
- Objectif : mesurer et encourager la recherche dans le traitement automatique de messages (objectif militaire, initialement)

⇒ évaluation régulière des avancées scientifiques

⇒ Mode d'évaluation ouvert et coopératif (comparaison des résultats et des technologies)

⇒ introduction des mesures de rappel/précision

Déroulement d'une évaluation MUC

- Un domaine est d'abord fixé par les évaluateurs
 - messages de la marine américaine
 - récits d'attentats terroristes en Amérique du Sud
 - messages concernant les joint-ventures
- étape de préparation (annotation des données, adaptation des systèmes)
- une étape d'évaluation par un système de boîte noire (comparaison automatique des résultats fournis par les systèmes avec ceux fournis par un expert)
- une étape d'échange entre concepteurs de systèmes

Bref historique des campagnes MUC

- MUCK-1, 1987 : messages de US Navy
- MUCK-2, 1989 : formulaire de dix champs à remplir, premières mesures d'évaluation
- MUC-3: 1991: 18 champs, récits d'attentats terroristes en Amérique du Sud, affinement des mesures d'évaluation
- MUC-4: 1992: 24 champs, corpus similaire à MUC-3

Bref historique des campagnes MUC

- MUC-5: 1993: 47 champs, en anglais et japonais, messages concernant des joint-ventures et de la microélectronique.
 - Première tentative d'améliorer la portabilité (multi-linguisme, multi-domaines)
- MUC-6: 1995: messages concernant les changements dans les postes de direction d'entreprises ; 4 tâches sont définies : reconnaissance d'entités nommées, co-référence, formulaire des entités, formulaire des scénarios.
 - définit des modules génériques et réutilisables par delà les domaines
- MUC-7: 1998 : dernière conférence
 - application des techniques d'apprentissage à l'extraction d'informations

Evolution

- Recherche d'informations de plus en plus précises
 - de 10 champs à remplir lors de MUC-2 (1987) à 47 pour MUC-5 et près de 100, structurés en formulaires, pour MUC-6 (1995)
- Définition de modules réutilisables génériques
 - Entités nommées
 - Analyse de la co-référence
 - Formulaire d'entité
 - Scénario
- Analyse locale généralement effectuée par un système de transducteurs appliqués en cascade
- Affinement des mesures d'évaluation

Les scores

- Rappel/Recall ($\neq$ surdité)

$$\text{Rappel} = R = \frac{\text{nb. réponses correctes}}{\text{nb. total de réponses attendues}}$$

- Précision ($\neq$ Jeanne d'Arc)

$$\text{Précision} = P = \frac{\text{nb. réponses correctes}}{\text{nb. de réponses trouvées par le système}}$$

- F-mesure

$$F - \text{mesure} = \frac{(\beta + 1) \times P \times R}{(\beta \times P) + R}$$

- Autres questions en rapport :

- correspondances partielles, champs plus ou moins importants, ...

Limitations des systèmes MUC

- Les résultats semblent avoir atteint un plafond
 - la plupart des systèmes étaient basés sur une analyse de surface à base de transducteurs à états finis.
- L'adaptation à un nouveau domaine restait difficile
 - La constitution des ressources est coûteuse en temps (systèmes à base de connaissance)
 - Comment articuler les connaissances générales et celles dépendant du domaine
 - Vers des techniques d'apprentissage
- Extraction d'Information multi-documents ?
 - Comment traiter une vue différente, des commentaires, de l'annotation d'une information précédente ?

1 Les conférences MUC

2 **Extraction d'informations avec des transducteurs à états finis**

Principaux problèmes de la compréhension de texte

- Qualité moyenne (Tacitus: F-mesure de 57 à MUC3)
- Bonne précision mais mauvais rappel
- Temps de traitement (36 heures pour 100 dépêches)
- Difficile à adapter à un nouveau domaine

Systèmes plus simples

- Ciblés sur de l'information pertinente
- Basés sur des transducteurs à états finis
- Exemple: Passage de Tacitus (Hobbs et al., 1991) vers Fastus (Appelt et al., 1993)

Document

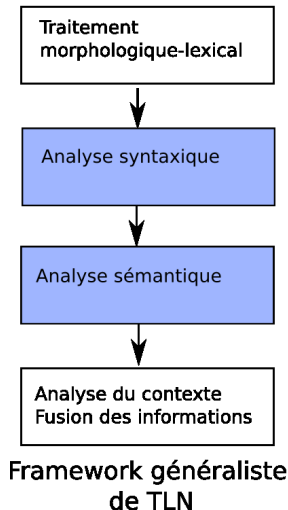
Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan. The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

Patron à remplir

- **Relation:** TIE-UP
- **Joint Venture:** “Bridgestone Sports Taiwan Co”.
- **Entités:** “Bridgestone Sport Co.”, “a local concern”, “a Japanese trading house”
- **Activité:** “iron and *metal wood* clubs”
- **Capital:** NT \$200000000, **Date de création:** During january 1990

Architecture

- Cascade de transducteurs à états finis
- Etapes:
 - Traitement lexical
 - reconnaissance des mots complexes
 - Traitements syntaxiques
 - reconnaissance des groupes nominaux simples
 - reconnaissance des groupes nominaux complexes
 - Traitement sémantique
 - reconnaissance des événements du domaine
 - Remplissage du patron d'extraction
 - fusion des informations



1. Mots complexes

Reconnaitances des expressions multi-mots et des noms propres

2. Détection des groupes nominaux simples

Identification des groupes nominaux, groupes verbaux

3. Détection des groupes nominaux complexes

Identification des groupes nominaux et groupes verbaux complexes

4. Évènement du domaine

Règles d'extraction des évènements intéressants pour le domaine, des patrons d'extraction sont construits

5. Fusion des informations

Les informations extraites du texte sont fusionnées si elles concernent la même entité/événement

- Reconnaissance des nombres
- Reconnaissance des noms propres (listes de prénoms, de lieux . . .)
- Reconnaissance des amorces comme M. (pour Monsieur) ou SA (pour Société anonyme)
- Reconnaissance et normalisation des sigles comme I.B.M ou I.B.M. (avec ou sans point à la fin)
- Étiquetage des mots inconnus et des mots commençant par une majuscule
- Utilisation d'Unitex

2 - Traitements syntaxiques : Multi-mots

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

- Entités nommées
- Expressions multi-mots

Principe

Bridgestone Sports Co./CompName

said/VG

Friday/NG

it/NG

had set up/VG

a joint venture/NG

in/PREP

Taiwain/Location

With/Prep

a local concern/NG

and/Conj

a Japanese trading house/NG

- Identification de groupes (nominaux, verbaux, ...)
- Pas de résolution des attachements

Exemple de grammaire d'extraction (UNITEX)

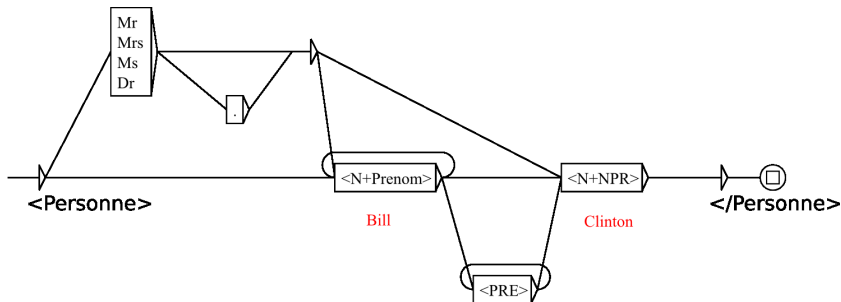


Figure: Détection des noms de personnes

Exemple de grammaire avec imbrication (UNITEX)

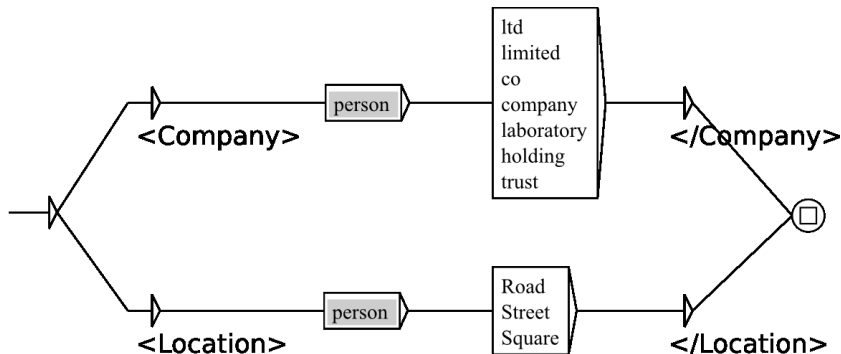


Figure: Extrait des noms de personnes entrants dans des noms d'entreprises ou de lieu

Principes

Les règles s'appliquent en respectant certaines heuristiques :

- Principe du “plus long patron” (longest match),
- Une règle ne peut plus s'appliquer à l'intérieur d'une séquence précédemment reconnue
- Si deux règles de même longueur peuvent s'appliquer, l'ordre d'application des grammaires joue le rôle d'arbitre

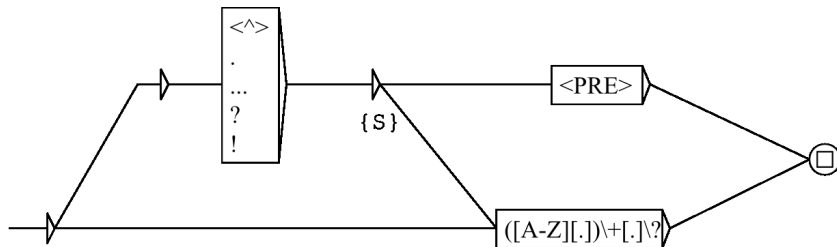


Figure: Exemple simple de détection de début de phrase

Application

Après de nombreux essais, la S.N.C.F. généralisa l'usage du T.I.A. en 1942. T.I.A. consiste à introduire...

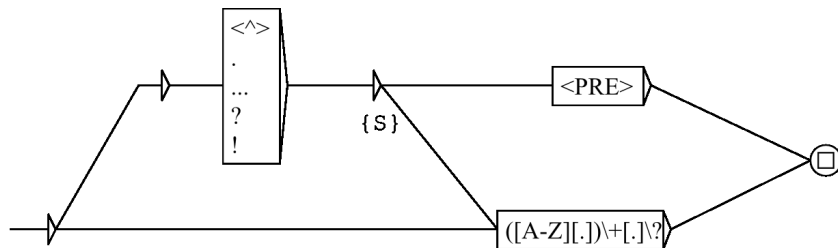
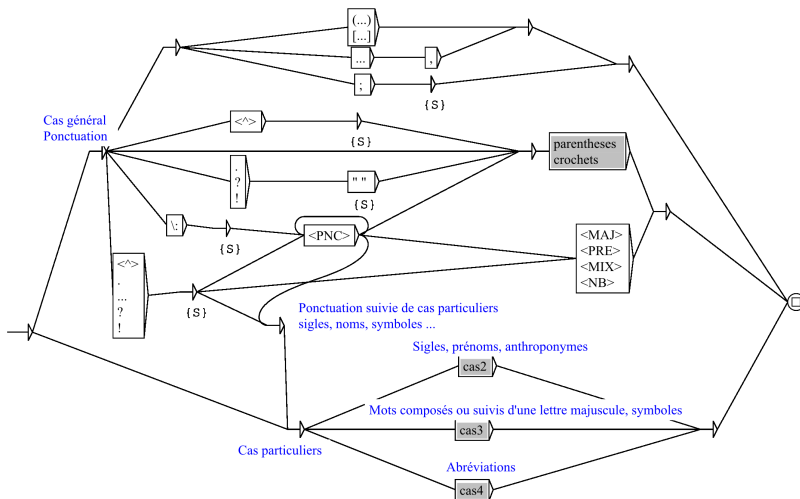


Figure: Exemple simple de détection de début de phrase

Application

{S}Après de nombreux essais, la S.N.C.F. généralisa l'usage du T.I.A. en 1942. {S}T.I.A. consiste à introduire...

Grammaire de détection de phrases



Améliorations proposées dans Fastus

Révision dynamique des étiquettes en fonction d'un contexte local (theory refinement, Mooney 1993, Dejean Coling 2000, Brill, ...)

- Par défaut, Washington est un nom de lieu
- D'après les contextes “Consuela Washington” et “Mrs. Washington”, retypage en noms de personnes des occurrences isolées.
- Bien adapté à des documents courts (domaine de localité)

Principe

- gestion de l'apposition, par exemple "Arizona senator Barry Goldwater won the Republican nomination", "Barry Goldwater" est un groupe nominal en apposition de "Arizona senator".
- construction de groupes indiquant une *mesure*
- conjonction entre groupes nominaux

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

Société

Produit

Argent

Date

Extraction d'évènements du domaine

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

Société

Joint Venture

Utilisation de patrons du type :

```
[COMPANY] [SET-UP] [JOINT-VENTURE] with [COMPANY]  
[COMPANY] [SET-UP] [JOINT-VENTURE] (others)* with [COMPANY]  
...
```

4 - Remplissage des patrons (Fusion des informations)

Texte

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

Patron

Relation: TIE-UP

Joint Venture: "Bridgestone Sports Taiwan Co."

Entités: "Bridgestone Sport Co."

"a local concern"

"a Japanese trading house"

Activité: "iron and 'metal wood' clubs"

Montant: NT\$200000000

Date: During January 1990

Performances des systèmes MUC

- Règles/Automates *faits mains* : jusqu'à 6 mois pour un seul domaine (MUC-6)
- Performances :
 - Repérage d'Entités Nommées : jusqu'à 0,9 de P&R.
 - Scénario : jusqu'à 0,65 de P&R (accord inter-annotateurs : entre 0,6 et 0,8 de P&R)
 - Est-ce suffisant en pratique ?
- Portabilité difficile (domaines, langages)
 - Approches par apprentissage automatique