ERNARD LYON-I
11 Novembre 1918
69621 VILLEURBANNE

Diplôme d'Etudes Supérieures Specialisées

Informatique documentaire

* ~~MEMOIRE DE STAGE~~
* NOTE DE SYNTHESE

OUTILS STATISTIQUES POUR LE TRAITEMENT AUTOMATIQUE DES LANGUES.

AUTEUR : VERBAERE Catherine

DATE : Juin 1980

INTRODUCTION

L'utilisation de l'analyse automatique des langues est une voie vers la solution des problèmes de traitement automatique de longs textes en langue naturelle.

Les méthodes linguistiques traditionnelles ne suffisent pas pour lever les ambiguités lors de l'analyse morphologique et de l'analyse syntaxique. L'analyse sémantique qui suit l'analyse syntaxique devrait éliminer ces ambiguités. Malheureusement l'analyse sémantique est difficilement concevable en traitement automatique des langues.

Depuis quelques années de nouvelles méthodes d'analyses sont apparues. Ce sont des méthodes probabilistes et statistiques. Elles permettent de lever certaines des ambiguités citées plus haut. Ces procédés statistiques ont été au départ utilisés pour l'indéxation de textes. Ils étaient constitués de simples comptages de mots dans le texte, de calculs de fréquences, de calculs de corrélation. Ces méthodes se sont considérablement développées utilisant la théorie de l'échantillonnage, les filtres statistiques, les fonctions discriminantes, les processus de Markov ou la méthode du maximum de vraisemblance.

Ces méthodes permettent d'obtenir de très bons résultats sur un univers de mots réduit. Ils le seront beaucoup moins dès qu'on élargit cet univers. On notera que ce problème est la plus grande difficulté du traitement automatique des langues.

Nous allons donc voir les différentes façons d'utiliser les outils statistiques dans le traitement automatique des langues : les premières pages de cet exposé seront consacrées à la recherche bibliographique faite en vue de cet exposé.

II RECHERCHE BIBLIOGRAPHIQUE.

Il est important d'expliquer de quelles façons a été abordée cette bibliographie. La base de cette recherche est un compte rendu de recherche de l'UER Informatique et Mathématiques en sciences sociales :

SIRAUCO Un système d'aide à l'étude morphosyntaxique de corpus - Par BOURGUIGNON HOLLARD et ROUAULT. [1]

On a pu de ce document retirer les mots clés qui semblaient intéressants - On a obtenu la liste suivante -

PROBABILITES
STATISTIQUES
ANALYSE MORPHOLOGIQUE.
ANALYSE SYNTAXIQUE.
ANALYSE MORPHOSYNTAXIQUE.
GRAMMAIRE.
INFORMATION RETRIEVAL.

Faisant un stage dans la société TELESYSTEMES QUESTEL : serveur français de bases de données j'ai profité de ma formation sur le logiciel MISTRAL. pour interroger le fichier PASCAL : fichier du CNRS -
Il faut préciser pour comprendre le listing

qui suit que dans le logiciel MISTRAL, les troncatures sont ? (1 ou 0 caractères) et + (nb de caractères illimité) et la procédure d'environnement alphabétique d'un terme est la procédure LE.

On a obtenu le listing suivant (voir page suivante). Il faut noter que parmi les 32 documents 10 seront retenus, les autres étant rejetés à cause de la langue ou du sujet trop éloigné du traitement automatique des langues.

Pour terminer complétement cette recherche. Il a fallu obtenir les documents primaires. Pour cela il existe le centre de documentation du CNRS A l'aide d'un code des periodiques recus dans ce centre [LOC-CNRS-15077 sur l'exemple page 7] on obtient le document.

On peut dire dès à present que ce sujet n'a pas fait l'objet d'une grande publication surtout en langue anglaise ou française. En effet lors de mon interrogation j'ai du supprimer environ $\frac{1}{2}$ des références qui étaient en langue russe allemande ou japonaise.

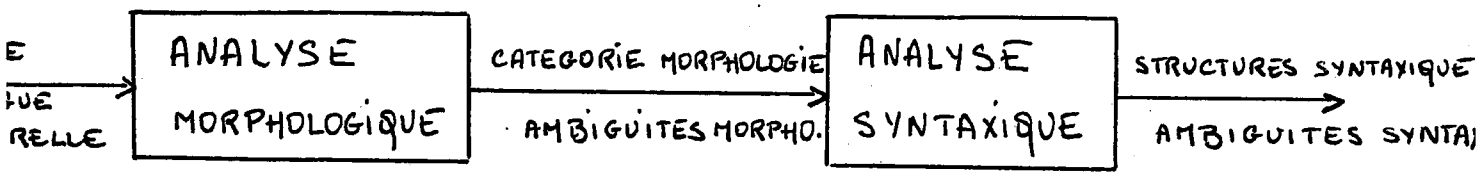
```
COM

??FRT0440804801401      V0204 P02
??    ST00        YOUR NAME IS CL02       DAY:0128,HOUR:0013,MIN:0058

??CENTSSTS     IS CONNECTED               DAY:0128,HOUR:0013,MIN:0058
PLEASE LOGIN: 0402,04201:89
PASSWORD: ****

LOGIN SUCCESSFUL. YOUR ID IS X287
  NOUVELLES SUR: BIBLIOS, PAPYRUS, IALINE.
!MISTRAL

  *QUESTEL*  mistral V5-00-01

  FRANCAIS, FRAPPER :1
  ENGLISH, TYPE :2
?1
  TERMINAL (T/VXX)
?T

  PROCEDURE: ..INFO, ..SOS, ..BASE ?

?..BA PASCAL

  BASE CONNECTEE: PASCAL
  PROCEDURE, OU ETAPE DE RECHERCHE  1

?STATISTIQUE?

  TERME MULTISENS STATISTIQUE?:  4
  T1    STATISTIQUE /DE
  T2    STATISTIQUE /UT
  T3    STATISTIQUES /DE
  T4    STATISTIQUES /UT
  SELECTIONNER OU NON ?

?S TT

  *1*  RESULTAT  25973
  PROCEDURE, OU ETAPE DE RECHERCHE  2

?PROBABILITE?

  TERME MULTISENS PROBABILITE?:  3
  T1    PROBABILITE /DE
  T2    PROBABILITE /UT
  T3    PROBABILITES /UT
  SELECTIONNER OU NON ?

?S TT

  *2*  RESULTAT   6658
  PROCEDURE, OU ETAPE DE RECHERCHE  3

?STOCHASTIQUE?

  TERME MULTISENS STOCHASTIQUE?:  2
  T1    STOCHASTIQUE /UT
  T2    STOCHASTIQUES /UT
  SELECTIONNER OU NON ?

?S TT

  *3*  RESULTAT   6030
```

```
?ANALYSE ET SYNTAXIQUE

  TERME MULTISENS ANALYSE:  2
  T1    ANALYSE /DE
  T2    ANALYSE /UT
  SELECTIONNER OU NON ?

?S TT

        RESULTAT 130236
  *4*   RESULTAT    740
  PROCEDURE, OU ETAPE DE RECHERCHE  5

?ANALYSE ET MORPHOLOGIQUE

  TERME MULTISENS ANALYSE:  2
  T1    ANALYSE /DE
  T2    ANALYSE /UT
  SELECTIONNER OU NON ?

?S TT

        RESULTAT 130236
  *5*   RESULTAT    240
  PROCEDURE, OU ETAPE DE RECHERCHE  6

?ANALYSE ET MORPHOSYNTAXIQUE

  TERME MULTISENS ANALYSE:  2
  T1    ANALYSE /DE
  T2    ANALYSE /UT
  SELECTIONNER OU NON ?

?S TT

        RESULTAT 130236
  *6*   RESULTAT      4
  PROCEDURE, OU ETAPE DE RECHERCHE  7

?GRAMMAIRE

  TERME MULTISENS GRAMMAIRE:  2
  T1    GRAMMAIRE /DE
  T2    GRAMMAIRE /UT
  SELECTIONNER OU NON ?

?S TT

  *7*   RESULTAT    565
  PROCEDURE, OU ETAPE DE RECHERCHE  8

?..HI

  ETAPE  FREQ
     1  25973 STATISTIQUE?
     2   6658 PROBABILITE?
     3   6030 STOCHASTIQUE?
     4    740 ANALYSE ET SYNTAXIQUE
     5    240 ANALYSE ET MORPHOLOGIQUE
     6      4 ANALYSE ET MORPHOSYNTAXIQUE
     7    565 GRAMMAIRE
```

```
  PROCEDURE, OU ETAPE DE RECHERCHE  9

?1 OU 2 OU 3

  *9*  RESULTAT  35863
  PROCEDURE, OU ETAPE DE RECHERCHE 10

?4 OU 5 OU 6 OU 7

  *10* RESULTAT   1211
  PROCEDURE, OU ETAPE DE RECHERCHE 11

?9 ET 10

  *11* RESULTAT     52
  PROCEDURE, OU ETAPE DE RECHERCHE 12

  PROCEDURE, OU ETAPE DE RECHERCHE 13

?ANALYSE MORPHOLOGIQUE

  *13* RESULTAT     50
  PROCEDURE, OU ETAPE DE RECHERCHE 14

?4 OU 6 OU 7 OU 13

  *14* RESULTAT   1025
  PROCEDURE, OU ETAPE DE RECHERCHE 15

?9 ET 14

  *15* RESULTAT     32
```

2684 C.PASCAL
: 78-3-0295447
: DELATTE L.; DUCHESNE-DEGEY M.; GOVAERTS S.; DENOOZ J.
: NIV. LIEGE, LAB. ANAL. STAT. LANGUES ANC., LIEGE, BELG.
: LE TRAITEMENT AUTOMATIQUE DE LA LANGUE FRANCAISE AU LABORATOIRE D'ANALYSE STATISTIQUE DES LANGUES ANCIENNES.
: TP;LM;DU
: ORGANIS. INTERNATION. ET. LANGUES ANC. ORDINAT., REV.; BELG.; DA. 1977; NO 4; PP. 1-55; BIBL. 6 REF.; LOC. CNRS-15077

I ANALYSE MORPHOLOGIQUE ET ANALYSE SYNTAXIQUE.



A - ANALYSE MORPHOLOGIQUE.

L'analyse morphologique est l'analyse des différentes unités lexicales composant le texte en langue naturelle [2] Une décomposition sera faite en éléments suivants : la base (ou racine), les affixes (préfixe ou suffixe) et la desinence.

PREFIXE	BASE	SUFFIXE	DESINENCE

On traite un texte en langue naturelle - la langue est un ensemble infini. où il existe un nombre de phénomènes différents - Ainsi on aura des difficultés dans

→ le choix de la base pour un terme

|IR|REMEDI|ABLE| base REMEDI

dans ce cas REMEDE sera analysé à part

→ Le choix de dislocation ou non d'expression

POMME DE TERRE.

→ le choix des catégories morphologiques

catégorie NOM. ou { catégorie NOM PROPRE / catégorie NOM COMMUN

On peut donc affiner ou non une catégorie à l'aide de variables.

Ainsi l'analyseur morphologique avec l'aide d'un dictionnaire de références où sont répertoriées : bases désinences, affixes, fera correspondre le terme à analyse avec un modèle suivant les règles morphologiques. Il y aura des ambiguités morphologiques.

B. ANALYSE SYNTAXIQUE

L'analyse syntaxique d'un texte en langue naturelle fonctionne avec une grammaire donnée à priori. L'ensemble des règles appartenant à la grammaire ne doit pas être trop grand, cela augmenterait les difficultés. Il faut donc analyser le plus grand nombre de phrase avec un nombre de règles correcte.

L'analyseur syntaxique valide une structure de phrase en fonction de la grammaire choisie au départ. Il pourra ainsi résoudre des ambiguités morphologiques. Il y aura cependant des ambiguités syntaxiques.

C- LES AMBIGUITES MORPHOSYNTAXIQUES.

Le traitement automatique des langues comporte une analyse morphologique et une analyse syntaxique - etant entendu que l'analyse semantique ne sera pas abordée ici.

Lors de l'analyse morphosyntaxique d'un texte plusieurs obstacles se présentent. Ces obstacles sont dus à la grande richesse de la langue naturelle. Ces ambiguités sont donc morphologiques [5]

Exemple : SI adverbe affirmation
adverbe intensité
conjonction de subordination

POUVOIR } substantif
ETRE } verbe.
AVOIR }

UN numéral.
pronom indéfini.
article.

Ces exemples montrent que lors de l'analyse morphologique d'une phrase le nombre de combinaisons possibles devient vite très grand.

Cependant l'analyse syntaxique éliminera beaucoup de ces ambiguités - Mais elle en générera

d'autres : des ambiguités syntaxiques -
Exemple : IL Y A UNE MAISON.

Cette phrase peut être analysée.
[IL Y A] [UNE MAISON]

Ou

[IL] [Y] [A] [UNE MAISON]
sous entendu à cet endroit.

Pour ce type d'ambiguités seule une étude du contexte de la phrase permet de savoir quelle est l'analyse correcte. Cet aspect est abordé normalement dans l'analyse sémantique.

IV OUTILS STATISTIQUES

Nous nous apercevons qu'une nouvelle approche du traitement automatique des langues est nécessaire, les outils classiques ne suffisent pas pour analyser automatiquement un texte.

On a été amené à introduire des méthodes statistiques - Un approfondissement de ces méthodes a permis à l'aide d'échantillon de déduire des propriétés linguistiques, des structures morphosyntaxiques d'un texte et d'éliminer la plupart des ambiguités, après les analyses morphologiques et syntaxiques. Ces méthodes sont quantitatives et permettent de prendre des décisions lors de chaque ambiguité réduisant ainsi au minimum le nombre de ces ambiguités.

Nous allons étudier ces différentes utilisations des statistiques et probabilités dans le traitement automatique des langues.
Nous verrons d'abord les simples comptages, les calculs de fréquences, les calculs de corrélation, puis les grammaires probabilistes, les filtres statistiques, enfin une approche de corrections d'erreurs syntaxiques.

I- PREMIERE APPROCHE.

Comme nous venons de le dire les premiers outils statistiques utilisés ont été les fréquences les coefficients de corrélation - L'inconvenient de ces méthodes est qu'elles ne peuvent être appliquées que sur un univers assez restreint - Dès qu'un texte est long ces calculs sont alors beaucoup trop longs et donc ils ne sont plus intéressants !

Nous verrons que ces méthodes sont utilisées pour construire des théories plus élaborées telles que les filtres statistiques - Elles sont cependant beaucoup exploitées en indexation automatique et en recherche d'évaluation de pertinence [3]

Nous donnerons quelques exemples de SALTON. sur la recherche d'evaluation de pertinence -

On pose $P(x_i/w_1)$ = probabilité qu'un terme x_i apparaissent dans un enregistrement donné et que cet enregistrement soit pertinent

$P(x_i/w_2)$ = probabilité qu'un terme x_i apparaissent dans un enregistrement donné et que cet enregistrement ne soit pas pertinent.

En utilisant la formule de Bayes la fonction de recherch

$$P(\omega_i|x) = \frac{P(x|\omega_i)\,P(\omega_i)}{\sum_{i=1}^{2} P(x|\omega_i)\,P(\omega_i)}$$

ou $P(\omega_i)$ est la probabilité a priori de pertinence
x est un vecteur de un ou plusieurs termes
En utilisant une approximation la fonction discriminante

$$g(x) = \frac{P(\omega_1|x)}{P(\omega_2|x)}$$

Le calcul pratique de $P(x|\omega_i)$ se fait en calculant la fréquence de x_i ou en utilisant un modèle de distribution de poisson -

Considérant le cas d'indépendance le calcul des fréquences se fait sur un échantillon. ($x_i = 1$ si x_i est un terme de x)

$$p(x|\omega_i) = \prod_{i=1}^{n} P(x_i=1|\omega_1)^{x_i} \left(1 - P(x_i=1|\omega_1)\right)^{1-x_i}$$

avec $p(x_i=1|\omega_1) = \frac{r_i}{R}$ $\quad p(x_i=1|\omega_2) = \frac{n_1 - r_i}{N-R}$

	ω_1	ω_2	
$x_i=1$	r_i	$n_i - r_i$	n_i
$x_i=0$	$R - r_i$	$N - n_i - R + r_i$	$N - n_i$
	R	$N-R$	N

table d'occurrence d'un terme x_i
et d'une question q.

1- GRAMMAIRES PROBABILISTES.

La théorie probabiliste a ensuite été utilisée à un niveau supérieur = au niveau des grammaires choisies lors de l'analyse syntaxique.

Rappel sur les grammaires [4]
Une grammaire formelle G est définie par le quadruplet

$$G = (V_T, V_N, R, A)$$

où V_T est l'ensemble des symboles terminaux
V_N est l'ensemble des symboles non terminaux
R est l'ensemble des règles
A est le symbole de départ.

Exemple $G = (V_T, V_N, R, S)$

$V_N = \{S, B, C\}$.
$V_T = \{a, b, c\}$.

R :
- $S \longrightarrow aSBC$
- $S \longrightarrow abC$
- $CB \longrightarrow BC$
- $bB \longrightarrow bb$
- $bC \longrightarrow bc$
- $cC \longrightarrow cc$

On définira une grammaire probabiliste à l'aide d'une grammaire formelle. Il existera en plus des probabilités associées à chaque règle de la grammaire. Une telle grammaire sera définie par un quintuplet [8]

Exemple.

$$G = (V_T, V_N, R, S, P)$$

P est l'ensemble des probabilités.
les autres symboles sont les mêmes que pour une grammaire formelle.

$V_N = \{\sigma, A_2\}$
$V_T = \{a, b\}$

R
$\sigma \longrightarrow a A_2 A_2$ ① P p_{11}
$\sigma \longrightarrow b$ ② p_{12}
$A_2 \longrightarrow A_2 \sigma$ ③ p_{21}
$A_2 \longrightarrow a$. ④ p_{22}.

$p_{11} = P(a A_2 A_2 / \sigma)$
$p_{12} = P(b / \sigma)$
$p_{21} = P(A_2 \sigma / A_2)$
$p_{22} = P(a / A_2)$

Si on a la chaine suivante $aaba = v$
v est obtenu

$$\sigma \xrightarrow{①} a A_2 A_2 \xrightarrow{③} a A_2 \sigma A_2 \xrightarrow{④} a a \sigma A_2 \xrightarrow{②} aab A_2 \xrightarrow{④} v$$

p(v) est obtenu

$$p(v) = P(a A_2 A_2 / \sigma)\ P(A_2 \sigma / A_2)\ P(a / A_2)\ P(b / \sigma)\ P(a / A_2)$$
$$= p_{11}\ p_{21}\ p_{22}\ p_{1,2}\ p_{2,2}.$$

Les propabilités associées à chaque règle d'une grammaire sont calculées en comptant le nombre moyen de fois que chaque règle est utilisée dans le langage généré par cette grammaire. Ce n'est pas toujours facile l'univers est trop grand. Aussi on travaille souvent sur un échantillon représentatif de ce langage.

Notons l'article de HARYANSKI qui traite un aspect un peu particulier des grammaires probabilistes [7]. En effet il traite des grammaires déterministes avec un langage fini. Après quelques propriétés sur ces grammaires particulières, il propose l'utilisation des processus de Markov.

Ces processus permettent de calculer la probabilité d'une chaine de caractères. Pour cela on doit d'abord calculer la matrice de transition du processus.

L'utilisation la plus intéressante qu'il fait de ces grammaires est la construction d'une méthode heuristique pour trouver une grammaire qui s'adaptera le mieux à un échantillon source.

Les outils statistiques nécessaires sont le test du χ^2. Ce test permettra de prendre la décision à propos de la grammaire proposée

En effet on pose

$$\chi^2 = \sum_{i=1}^{n} \left(f_i - M p(x_i)\right)^2 / M p(x_i)$$

avec S l'échantillon tel que $S = \{(x_1, f_1) \dots (x_n, f_n)\}$

tel que $\sum_{i=1}^{n} f_i = M$. $x_i \in L(G)$ f_i sa fréquence

et tel que la probabilité d'occurrence d'un échantillon S donné est.

$$P(S) = \prod_{i=1}^{n} p(x_i)^{f_i}$$

où $p(x_i)$ est la probabilité de la chaine x_i dans le langage $L(G)$.

Pour le test χ^2 on est obligé de définir un seuil α.

Ainsi la règle de décision sera

- Si $\chi^2 \geq \chi^2_\alpha$ on rejete l'hypothese que G est une grammaire correcte
- Si $\chi^2 < \chi^2_\alpha$ on accepte l'hypothèse que G est une grammaire correcte.

Il est important de calculer les probabilités des règles de la grammaire G.

Ce calcul que nous ne développerons pas ici est en fait le calcul de l'estimation par la méthode du maximum de vraisemblance.

On notera seulement le résultat

$$p_i = \frac{n_i}{n(A_i)}.$$

avec n_i : nombre de fois que la régle r_i avec la prémis A_i

$n(A_i)$ nombre de fois que le symbole non terminal A_i apparait dans les chaines de S.

Ainsi l'auteur nous propose l'algorithme suivant. Nous le donnerons sous forme très schematique pour définir plus clairement l'idée principale. Puis nous le détaillerons de façon plus mathématiques pour montrer à quels niveaux se trouverout les tests du χ_2

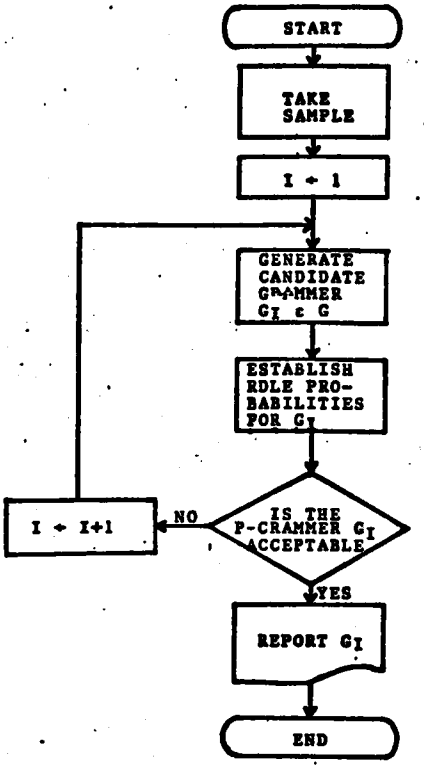


Fig. 1. General inference algorithm.

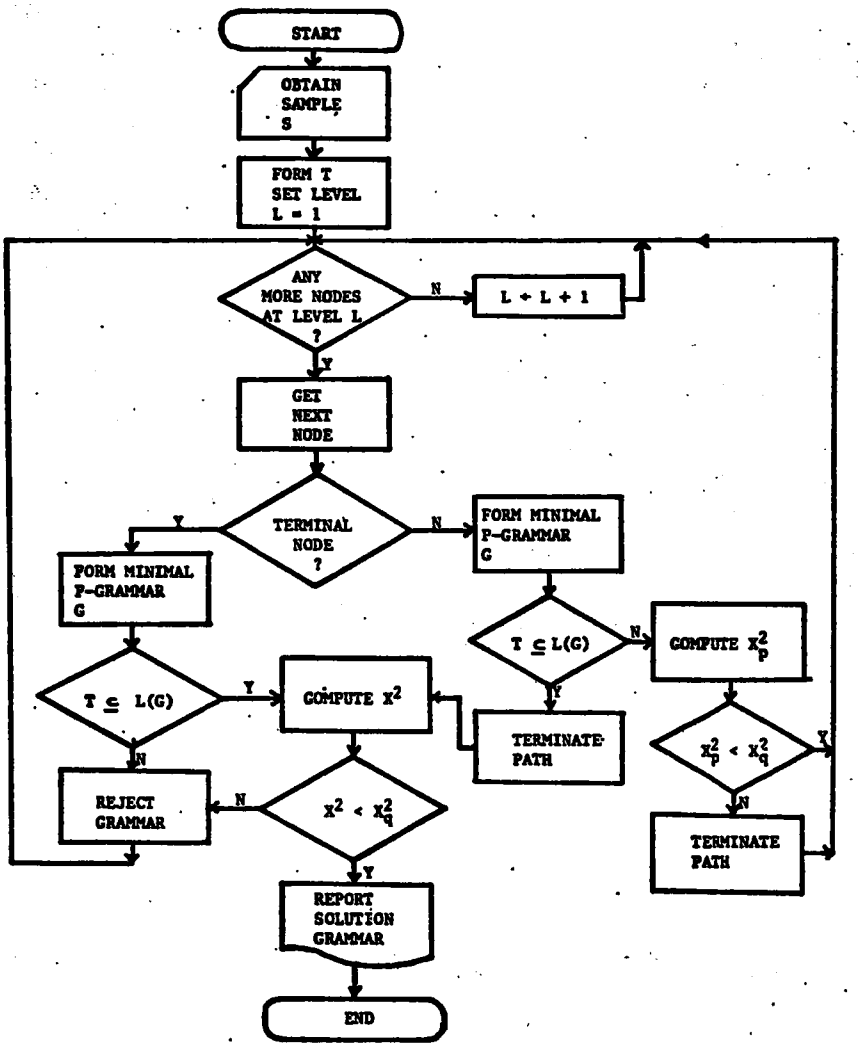


Fig. 3. Detailed inference algorithm.

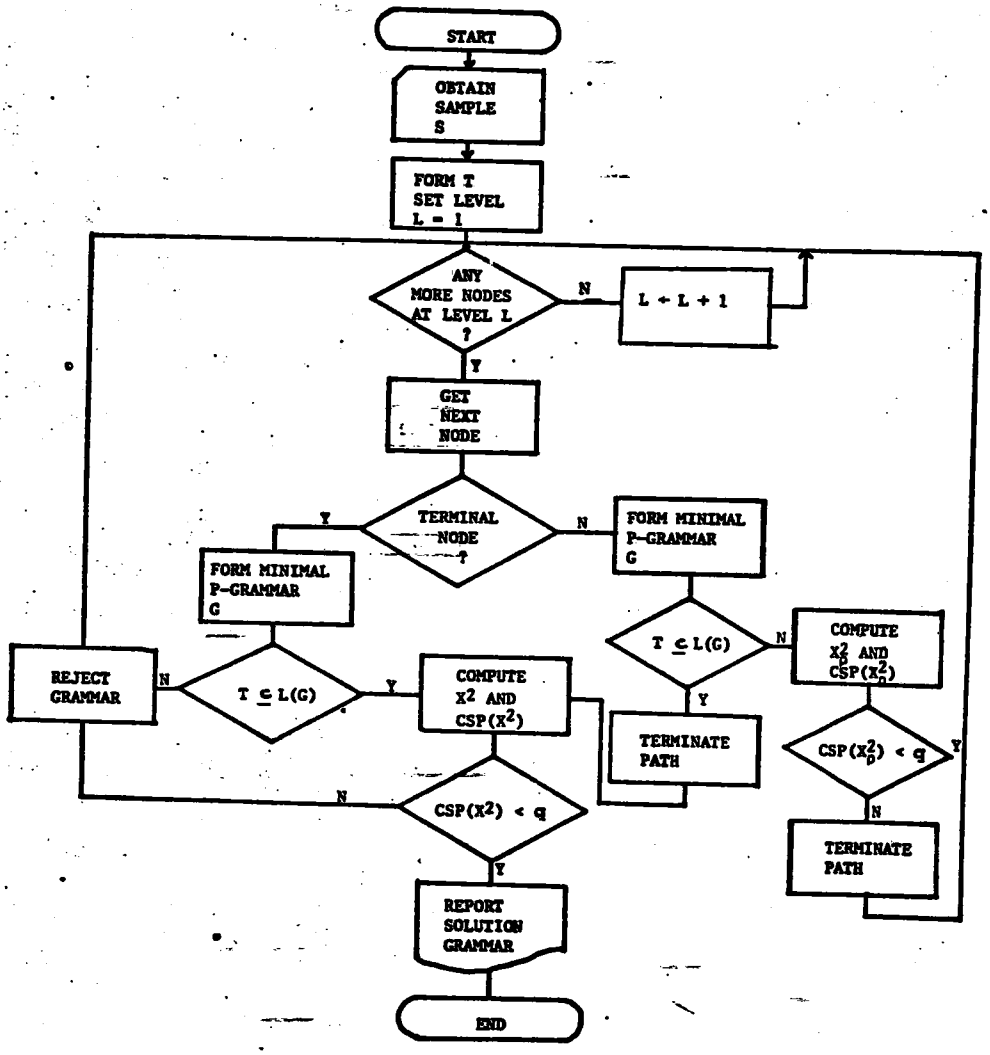


Fig. 4. Heuristic inference algorithm.

2. FILTRES STATISTIQUES

les grammaires probabilistes apportent des améliorations à l'analyse syntaxique. Les probabilités associées aux règles de grammaire permet une meilleure adaptation à un texte donné.

On a introduit en plus des grammaires probabilistes des filtres statistiques. Ce filtre permet, à la sortie de l'analyse morphologique, de supprimer la plupart des ambiguités [2]

le principe du filtre est construit sur les cooccurences des catégories morphologiques.
Ainsi on prend un échantillon de phrases tirées de façon aléatoire. Cet échantillon est analysé manuellement pour qu'il n'y ait plus d'ambiguité.
On calcule alors la matrice de effectifs de tous les couples de catégories morphologique
On obtient la matrice des fréquences qui permettra d'étudier la dépendance de deux catégories morphologiques. Ceci se fait a l'aide d'un test en χ^2 et de la table des contingences.
On définira aussi un test pour la table des effectifs pour évaluer la représentativité de l'échantillon.

Tous ces calculs statistiques établis sur l'échantillon forment la base du filtre -

Soit une phrase

$M_1(C_1)\ M_2(C_2) \quad M_n(C_n)$

M_i est un mot

C_i une catégorie morphologique

Après l'analyse morphologique on aura par exemple

$A(a_1 \ldots a_n)\ B(b_1 \ldots b_n)\ C(c_1 \ldots c_p)\ D(d_1 \ldots d_q)\ E(e_1 \ldots e$

On examine AB.

1. Si $n=1$ $m=1$ → pas d'ambiguité
2. Si $n \neq 1$ $m \neq 1$ → on utilise les résultats statistiques.

Après décision

3. Si $n=1$ $m=1$ → plus d'ambiguité
4. Si $n \neq 1$ $m \neq 1$ → on ne peut rien dire sur A
5. Si $n=1$ et $m \neq 1$ ou $n \neq 1$ et $m=1$ → on observe BC comme en 2

Après décision

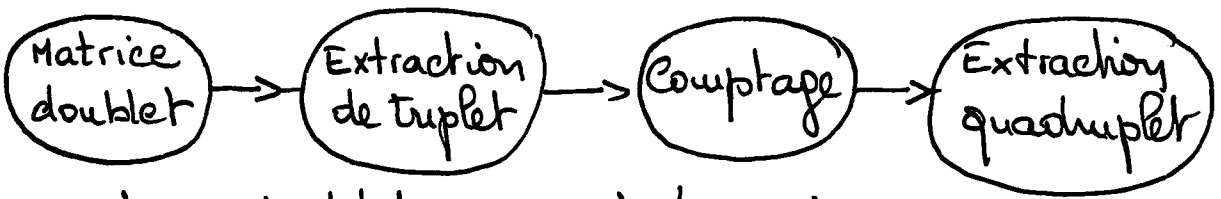
- Si $m=1$ → on reexamine $A(a_1 \ldots a_n)\ B(b_1)$ on passe à CD comme en 1
- Si $m \neq 1$ → on ne decide rien sur A et B on passe à CD comme en 1

Cet algorithme a l'avantage de ne pas être de taille très grande et minimise ainsi le temps -

les résultats expérimentaux montrent que sur un échantillon de 1491 mots sur un texte de 7000 mots représentant environ 40% de mots comportant au moins une ambiguité soit 3000 mots ambigus.

On obtient après le filtre
250 mots ambigus soit moins de 10%
60 mots comportant une erreur soit moins de 2%.

L'analyse syntaxique se fait par un système de recherche de suites de catégories morphologiques autres que les couples.
On propose ainsi de traiter les triplets (quadruplets, quintuplets) consécutifs et significatifs.
On calcule les fréquences relatives correspondante sur l'échantillon. Et on choisit à l'aide du test du χ^2 les triplets significatifs et ainsi de suite.



exemple doublet (DN,N) (N,AN)
triplet (DN,N,AN)

La construction du noyau grammatical se fait de façon dynamique - la grammaire évolue suivant des classes du corpus -

Monsieur HYA HYA utilise aussi la théorie du filtre statistique pour l'analyse morphologique [9] [10]
En effet on essaie d'obtenir automatiquement à l'aide de deux filtres statistiques les chaines initiales et terminales pertinentes en vue de l'analyse.
On veut obtenir un compromis entre la taille minimum des dictionnaires et le maximum de bruit morphologique.

Exemple la chaine terminale TIONS

TIONS chaine terminale de (verbe, substantif)

NOUS CHANTIONS

(Pronom) (Verbe substantif)

↓

bruit morphologique

La méthode consiste à effectuer un traitement statistique sur le texte d'entrée pour obtenir les différentes chaines et informations grammaticales et statistiques associés.
Ces mots non vides généreront des chaines terminale et initiales et ces chaines seront candidates pour être selectionnées chaines pertinentes

Plusieurs critères sont nécessaires pour établir ces deux filtres statistiques.

D'abord un critère de non redondance - Une chaine est non redondante par rapport à une autre chaine si elle n'apporte pas d'information supplementaire.

On définit ensuite les chaines primaires qui seront des chaines suffisamment longues et représentatives - On aura première espece et deuxieme espece.

Enfin on décrit un critère de pertinence
- pertinence morphologique
- pertinence grammaticale.

Le critère de pertinence morphologique minimise les chaines représentatives
le critère de pertinence grammaticale fixe le bruit morphologique tolérable.

Le premier filtre permet alors de selectionner sur les chaines primaires les chaines qui verifient
-> le critère de pertinence grammaticale
-> le critère de pertinence morphologique
-> le critère de non redondance

le deuxieme filtre après un tri sur les chaines du premier filtre traite en faite les chaines primaires dites de deuxième espece.

On peut après ces deux filtrages passer a

l'analyse morphologique - On a généré le système avec la consultation de différents lexiques - La consultation de ces lexiques se fait par ordre de priorité -

Le taux de résolution est de l'ordre de 56 %
Les resultats experimentaux montrent que la différence des taux de resolution entre l'apprentissage et les textes nouveaux diminue lorsque la longueur des textes augmente (ordre de grandeur 6 à 3 %).
Cette différence est du essentiellement au mot inconnu employé exceptionnellement.

D- CORRECTIONS D'ERREURS SYNTAXIQUES

Les ambiguités ne sont pas les seuls problèmes de l'analyse syntaxique - les erreurs en sont d'autre tout aussi importantes - les méthodes de corrections d'erreurs syntaxiques ont été largement développées avec la compilation principalement [11]

Monsieur Fu introduit pour les erreurs de l'analyse syntaxique un modèle statistique en vue de la corrections de ces erreurs -
Il distingue 3 types d'erreurs -

- substitution T_S
- suppression T_D
- insertion T_I .

Soit L un langage et Σ un alphabet fini, une chaine de L peut être déformée en une chaine différente n'appartenant pas a ce langage -

$x\,b\,y \in T_S(x\,a\,y)$ $\quad b \neq a$
$x\,a\,y \in T_I(x\,y)$ $\quad a$ dans Σ
$x\,y \in T_D(x\,a\,y)$ $\quad a$ dans Σ

On peut alors associer à chaque cas une probabilité, définissant ainsi le modèle stochastique.

la probabilité de substituer a par b est $q_S(b|a)$
la probabilité d'insérer b avant a est $q_I(b|a)$
la probabilité de supprimer a dans une chaine est $q_D(a)$
la probabilité d'insérer a a la fin d'une chaine est $q'_I(a)$

On calculera alors la probabilité qu'un symbole a soit déformé en α

$$q(\alpha|a) = q_D(a) \quad \text{si } \alpha = \Lambda$$
$$= \max(q(a), q_I(a|a)\, q_D(a)) \quad \text{si } \alpha = a$$
$$= \max(q_S(b|a), q_I(b|a)\, q_D(a)) \quad \text{si } \alpha = b \neq a$$
$$= q_I(b_1|a) \cdots q_I(b_{l-1}|a) \max\{q_S(b_l|a), q_I(b_l|a) q(a)\} \quad \text{si } \alpha = b_1 \cdots b_l \; b_l \neq a \; l > 1$$
$$= q_I(b_1|a) \cdots q_I(b_{l-1}|a) \max\{q(a), q_I(b_l|a)\, q_D(a)\} \quad \text{si } \alpha = b_1 \cdots b_l \; b_l = a \; l > 1$$

La méthode utilisée pour essayer de corriger ces erreurs est un algorithme utilisant la théorie du maximum de vraisemblance.
Elle est définie ainsi y n'appartenant pas au langage $L(G)$ et x appartenant a $L(G)$

$$q(y|x)\, p(x) = \max_z \{q(y|z)\, p(z) \mid z \in L(G)\}.$$

où $p(z)$ est la probabilité de générer z avec G grammaire probabiliste -

Une telle méthode est utilisée sur des chaines de caractères - Ce n'est pas en relation directe avec

le qui nous interesse. Il est cependant important de noter la méthode qui par certain aspect pourrait être exploités dans de nouvelles recherches sur l'analyse syntaxique.

V CONCLUSION

le traitement automatique des langues est utilisé dans différents projets : la traduction automatique, l'analyse de contenu, la recherche de l'information.

Des problèmes théoriques du traitement automatique ont amené les chercheurs à utiliser les théories statistiques pour palier ses difficultés - L'analyse morphologique et l'analyse syntaxique se trouvent améliorées par des filtres statistiques, des grammaires probabiliste et la construction d'algorithme. Il est à remarqué que le temps de traitement des algorithmes est de plus en plus court.

L'évolution la plus intéressante est la génération de noyaux de grammaire qui peuvent ne pas rester statiques.

Cependant il est un inconvénient majeur, qu'on a retrouvé tout le long de l'exposé. Le traitement automatique des langues est très difficile à réaliser lorsqu'on travaille sur des texte très longs.

- BIBLIOGRAPHIE .

1 BOURGUIGNON C., HOLLARD S., ROUAULT J. ; SIRAUCO UN SYSTEME D'AIDE A L'ETUDE MORPHO-SYNTAXIQUE DE CORPUS ; DOC UER INFORMATIQUE ET MATHEMATIQUE EN SCIENCES SOCIALES ; DOC n°5 SERIE C ; UNIVERSITE SCIENCES SOCIALES GRENOBLE .

2 CARRETERO J. ; MESURES DE PROPRIETES STATISTIQUES DE TEXTE EN LANGUE NATURELLE POUR LA DEFINITION DE MODELES LINGUISTIQUES UTILISABLES EN ANALYSE MORPHO-SYNTAXIQUE ; THESE DOCTORAT ES SCIENCES GRENOBLE ; OCTOBRE 77 .

3 SALTON G. ; MATHEMATICS AND INFORMATION RETRIEVAL ; J. DOCUMENT. ; 1979 ; V 35 ; 1-29 ; 71 BIBL .

4 TAYLOR L. BOOTH , RICHARD A. THOMPSON ; APPLYING PROBABILITY MEASURES TO ABSTRACT LANGUAGES ; IEEE TRANS. COMP. ; 1973 ; V C-22 ; N 5 ; 442-450 .

5 DELATTE L., DUCHESNE-DEGEY M. , GOVAERTS S., DENOOZ J. . LE TRAITEMENT AUTOMATIQUE DE LA LANGUE FRANÇAISE AU LABORATOIRE D'ANALYSE STATISTIQUE DES LANGUES ANCIENNES ; ORGANIS. INTERN. ET. LANGUES ANC. ORDINATEUR ; 1977 ; N 4 ; 1-55 ; 6 BIB

6 FU K. S. ; SOME APPLICATIONS OF STOCHASTIC LANGUAGES ; IN. APPL. STAT. SYMP. PROC. , DAYTON OHIO ; 1976 ; 183-200 ; 1P BIBL.

7 MARYANSKI F. J., BOOTH T. L. ; INFERENCE OF FINITE-STATE PROBABILISTIC GRAMMARS ; I.E.E.E. TRANS. COMP. ; 1977 ; V 26 ; N 6 521-536 ; 12 BIBL .

8 VAN DER MUDE A., WALKER A. ; ON THE INFERENCE OF STOCHASTIC REGULAR GRAMMARS ; INFORM. AND CONTROL ; 1978 ; V 38 ; N 3 ; 310-329 ; 19 BIBL .

9. HLAL YAHYA ; ANALYSEUR MORPHOSYNTAXIQUE ET SYNTAXIQUE DE LANGUES NATURELLES OBTENU PAR APPRENTISSAGE ; NOTE CEA N 2053 ; 1978 ; 2 FASC ; 64 + 47 ; 106 BIBL.

10. HLAL YAHYA : METHODES D'APPRENTISSAGE POUR L'ANALYSE MORPHOSYNTAXIQUE ; THESE SCI. PHYS. INTELLIGENCE ARTIFICIELLE LINGUIST. AUTOM. PARIS ; 1979 ; (8) - 285 ; 111 BIBL.

11 SHIN-YEE LU , KING-SUN FU : STOCHASTIC ERROR-CORRECTING SYNTAX ANALYSIS FOR RECOGNITION OF NOISY PATTERNS ; I.E.E.E. TRANS. COMPUTERS. ; 1977 ; V 26 ; N 12 ; 1268 - 1276 ; 17 BIBL.