STUDIA UNIVERSITATIS BABEŞ-BOLYAI, PHILOLOGIA, LIII, 3, 2008

METHODES ET ALGORITHMES DANS LES PROCESSUS DE CATEGORISATION ET D'EXTRACTION DES INFORMATIONS

MANUELA MIHĂESCU[*], SANDA CHERATA[**]

ABSTRACT. Data (texts) structure and formats are major factors in their handling and querying. Processing the information contained in various texts implies performing actions on documents parameters, as well as using a controlled language. Once created, the documents may be subjects of automatic procedures based on linguistic processes such as: classifying/categorization, automatic summarization, information extraction, information retrieval, machine translation, etc. This paper presents some aspects of natural language processing and modeling implied in applications of great interest in documents processing: searching/ retrieval, information extraction and categorization.

Keywords: natural language processing, classifications, information retrieval

Les nouvelles technologies, notamment celles qui sont impliquées dans la communication, ont permis l'approche d'une grande quantité d'information, cependant que l'accès à celle-ci suppose de moins en moins de difficultés pour l'utilisateur. Cela étant, une nouvelle perspective sur les méthodes de traitement a dû être envisagée.

Dans les processus de traitement de l'information les signes, surtout les signes linguistiques (écrits ou oraux), ont un rôle principal parce que, par leur biais, «l'information devient objet d'analyse» (Golu, 1975: 160) et, par leur combinaison, on peut rendre, plus facilement, le contenu quantitatif et qualitatif de celle-ci. De plus, le développement d'Internet et des technologies web a mis à la disposition des utilisateurs une vaste quantité d'information, la plupart sous forme de signes linguistiques organisés dans des textes/documents[1].

Grâce aux recherches les plus récentes dans le domaine de traitement du langage naturel, on peut réaliser, sur les documents, plusieurs opérations comme: l'annotation, l'extraction d'informations pertinentes et la création des bases de connaissances structurées (l'ingénierie de la connaissance). De vraies technologies de recherche/découverte d'information, d'extraction et de catégorisation de celle-ci se sont développées dans un intervalle de temps en somme assez court.

[*] *Babes-Bolyai University, Cluj-Napoca, Romania*. E-mail: manuela_mihaescu@yahoo.com
[**] *Babes-Bolyai University, Cluj-Napoca, Romania*. E-mail: sanda_cherata@yahoo.com

[1] Le terme *texte* désigne ici un ensemble de données plus ou moins structurées, cependant que le terme *document* s'applique à un ensemble de données structurées.

Les techniques utilisées dans ces processus se basent non seulement sur des modèles des grammaires génératives (grammaire des constituants, grammaire transformationnelle, grammaire d'unification, etc.), mais aussi sur des modèles mathématiques (surtout le modèle de la chaîne Markov et le modèle Markov caché), dans lesquels la structure des séquences s'établit, à partir d'un état initial, - et en appliquant des règles et des contraintes, par des choix successifs (le cas de l'analyse syntaxique), ou, par une analyse structurelle (analyse morphologique). Lors de la recherche d'information on utilise fréquemment aussi les méthodes statistiques qui sont très efficaces (si l'on parle de grands corpus de textes).

Le langage, représentation et concepts

Considéré comme un système de signes et de règles, le langage peut être analysé autant au niveau du signe de codification (niveau informationnel - statistique), qu'au niveau des règles de combinaison de ces signes, par les relations syntaxiques, sémantiques et pragmatiques (niveau relationnel). En accord avec les théories linguistiques, le langage humain peut être caractérisé (Chomsky, 1969, 1996; Vasiliu, 1970 ; Saussure, 1998) par:

- les signes (unité de base) qui constituent le langage et qui ont un caractère arbitraire; ils sont des signes de nature symbolique (caractère arbitraire du lien entre signifiant/signifié);
- le mode de combinaison/recombinaison entre ces signes (l'aspect syntaxique et morphologique), très important pour la construction du sens;
- le caractère relationnel des signes linguistiques; la plupart des mots sont polysémiques;
- des règles de composition (de ces unités) par lesquelles les mots s'organisent en proposition; le nombre de propositions correctes (ou acceptées/acceptable) est théoriquement infini.

À partir de ces principes, les recherches sur le traitement du langage peuvent avoir deux perspectives (Jackson, Moulinier, 2002): la première est basée sur les théories linguistiques traditionnelles qui étudient le langage employant les modèles des grammaires génératives et transformationnelles (recherche sur les règles par lesquelles les signes se combinent pour construire des énoncés). L'autre perspective est celle des analyses statistiques qui établissent des règles en retrouvant des modèles (reconnaissance des formes) à partir des probabilités des occurrences, utilisant de grands corpus. Cette méthode est considérée, parfois, comme une méthode empirique.

Conformément aux théories formelles, on peut analyser le langage par les signes linguistiques dont il est constitué. La modélisation de ces signes est faite par la grammaire, «l'instrument formel qui définit la relation entre le contenu et le sens d'une phrase», par des analyses syntaxiques et sémantiques, c'est-à-dire, par la combinaison et le calcul des interprétations. Dans la plupart des cas, ces règles sont traitées indépendamment d'éléments à traiter. Les modèles qui sont créés peuvent

ainsi être appliqués au langage naturel. De plus, comme la linguistique formelle ne peut pas expliquer, à elle seule, la structure du langage, des *techniques d'inférences* (Tătar, 2003) ont été incorporées pour interpréter (en fait *calculer*) la signification des représentations. Ces techniques sont utilisées en relation avec les déductions logiques, la logique des prédicats, des méthodes de réduction, les formalismes sémantiques, etc.

Le traitement du langage naturel réfère aux technologies (logiciels et matériels) utilisées par des ordinateurs pour l'analyse et la synthèse du langage humain écrit et parlé. Si pour les langages artificiels, qui ne posent pas de problèmes d'ambiguïté ou d'influence des facteurs extralinguistiques, on peut créer facilement des grammaires formelles, dans le cas du langage naturel la situation est plus compliquée. Là on peut parler d'ambiguïté à tous les niveaux (lexico-morphologique, syntaxique et sémantique), mais aussi de dépendance du contexte, de *subjectivité*, d'*intentionnalité*. C'est pour cela que, dans les nouvelles recherches, on prend en considération tous ces facteurs en combinant les modèles génératifs avec les modèles mathématiques.

Les modèles mathématiques

Ces modèles[2] utilisés pour le traitement du langage sont basés, en général, sur des méthodes statistiques, ayant une grande efficacité, particulièrement pour l'étude du langage parlé (la synthèse ou la reconnaissance de la parole). Dans le domaine du langage écrit, ces modèles sont développés à la suite de l'apparition des corpus (de grands ensembles de textes) numérisés.

L'un des modèles les plus connus est le *Modèle Markov*[3] qui est utilisé, en général, pour la description de l'état d'un système dont la future évolution dépend des états précédents.

Dans les analyses grammaticales on utilise fréquemment le *Modèle Markov caché*, (MMC) ou HMM – (*Hidden Markov Model*); celui-ci a un grand degré d'abstraction, incluant une structure cachée (avec des paramètres inconnus) qui peut être déterminée «par les paramètres observables». Le modèle est utilisé en général pour mesurer la probabilité qu'un système se trouve dans un certain état, la durée de cet état et aussi la probabilité de transition d'un état à l'autre. Dans les cas d'analyses linguistiques, le modèle Markov caché est utilisé pour la représentation de l'unité linguistique et le calcul de la probabilité d'une séquence particulière (par l'algorithme de Viterbi), aussi bien que pour retrouver l'ensemble d'états le plus probable et les probabilités des sorties pour chaque état (Manning, Schütze, 1999).

[2] «Les modèles mathématiques de la langue sont des constructions qui retiennent certains aspects relationnels des phénomènes linguistiques» (Vraciu, 1980: 323).

[3] Le processus Markov est un processus aléatoire, dynamique, qui décrit l'état d'un système dont l'évolution future dépend strictement de l'état antérieur (Manning, Schütze, 1999).

Le Modèle Markov dans l'analyse syntaxique

«Syntax is the structure of the visible (or audible) forms of language» (Winograd, 1987). Dans sa définition la plus simple la syntaxe pourrait être considérée comme une analyse des mots (les constituants de base), et notamment, de la succession des mots (Cole, 1996), ou bien comme «un algorithme qui sait générer des séries d'éléments acceptables ou non-acceptables» (Eco, 2007: 236). On peut affirmer que pour être acceptables (corrects) certains mots sont conditionnés par d'autres. Une modalité de quantifier cette règle est celle d'utiliser une liste de *n-gramme*[4] qui est simplement une succession de *n* éléments. Si on prend n=2, on obtient des *bi-grammes*. Le plus souvent utilisé, pour les strategies de recherche des propositions correctes du point de vue lexical, c'est le *tri-gramme* (n=3) (Cole, 1996). Le principe de *n-gramme* est que le *n*-ième élément d'une suite ne dépend que des *n-1* ièmes qui le précèdent. Ainsi, dans le traitement d'un texte, si on identifie un certain mot, on dispose, grâce aux *bi-grammes*, par exemple, de la probabilité d'apparition du mot qui le suit immédiatement.

En utilisant ces méthodes de probabilités et de co-occurrence des mots pour des grands corpus, on peut retrouver la modalité d'organisation des mots en proposition. L'avantage de cette méthode est qu'elle couvre une grande diversité de langages en acceptant un vocabulaire étendu.

Une analyse syntaxique complète ne se réduit certainement pas à la modélisation de la succession des mots (Cole, 1996). La linguistique moderne considère que la syntaxe est une partie de la cognition (qui inclut aussi des processus cognitifs), et que, par ailleurs, la structure grammaticale est une «prémisse de la signification». Le modèle *n-gramme* n'est pas suffisant pour une analyse complète. Ainsi, intégre-t-on dans l'analyse syntaxique: le traitement des classes grammaticales (catégorie qui précise la fonction du mot dans la proposition), des règles de combinaisons, mais aussi, des méthodes de probabilité et stochastiques (Brown, P.F, et coll., 1992 ; Perraud, F., et coll., 2003). Le modèle MMC est fréquemment utilisé dans le système de reconnaissance de la parole (Manning, Schütze, 1999 ; Gales, Young, 2008), de reconnaissance des écritures manuscrites (OCR - *Optical Character Recongnition*), de même que dans les situations telles: la catégorisation, la transcription lexicale, la reconstruction de texte, etc.

Les modèles utilisés dans l'analyse sémantique

«S*emantics is the systematic relation between structures in a language and a space of potential meanings»* (Winograd, 1987). En ce qui concerne la sémantique du langage naturel il y a des recherches qui étudient la signification, la relation entre le mot et le sens du mot. L'objectif de ces études est de faire de sorte

[4] Le *N-grame* constitue une sous-séquence de n-caractères, d'une suite de caractères (non espacés) (http://en.wikipedia.org). Le modèle est utilisé à large échelle dans les analyses statistiques du langage.

que les ordinateurs puissent déceler la signification correcte d'un mot, d'une proposition ou d'un discours.

À la différence de la syntaxe qui comprend la spécification d'un alphabet et des règles pour construire des énoncés corrects, la sémantique d'un langage peut être définie comme un «modèle» où elle peut être exprimée mathématiquement par une fonction qui «donne une valeur à une expression» (Moechler, Auchlin, 2005).

Montague affirme, par ailleurs, qu'il existe des ressemblances entre le langage naturel et le langage formel, dans le sens qu'on peut appliquer le principe de compositionnalité sémantique[5] pour le deux types de langages.

La description sémantique du langage peut être associée à un automate à l'état fini qui permet de préciser les conditions de «vérité» pour chaque élément du langage. L'analyse des propositions plus complexes peut être faite à partir des constituants de base que sont les mots et les morphèmes.

Les modèles de combinaison des constituants sont déterminés par la structure syntaxique. Il y a plusieurs applications qui sont basées sur les théories de Montague, qui associent pour chaque expression deux caractéristiques: l'expression et le sens de l'expression dans la phrase d'un côté, et de l'autre, une valeur de «vérité» d'une phrase (intentionnalité et extensionnalité). La sémantique du langage naturel (à la différence des langages artificiels) «n'est pas isomorphe avec la syntaxe», ce qui lui confère une certaine ambiguïté. Une phrase ambiguë, c'est une phrase dont la structure de surface est le produit d'au moins deux structures profondes» (Vasiliu, Golopenția, 1969; Moechler, Auchlin, 2005: 114). Les ambiguïtés peuvent donc apparaître à tous ces niveaux.

Pour la résolution des ambiguïtés lexico-morphologiques on emploie les méthodologies de désambiguïsation du sens du mot (*Word Sense Disambiguisation*) qui incorporent des techniques de catégorisation, et plus récemment, des techniques d'apprentissage inductif.

Le problème de désambiguïsation lexicale comporte différentes situations (Tătar, 2003 ; Mihalcea, Pedersen, 2005; Jurafsky, Martin, 2008). La plus simple est celle qui se réfère aux mots polysémiques, cas où on utilise des techniques de classification. L'algorithme employé pour la classification des mots polysémiques (verbes ou substantifs) calcule la probabilité d'apparition d'un mot dans un contexte considéré «le plus important».

Il y a des situations où l'apparition d'un mot est conditionnée par la présence d'un autre; par exemple dans un document structuré, le sens d'un mot est, en général, le même dans le document entier. Dans ce cas on peut facilement introduire des contraintes et des règles de combinaison. Un autre type de situation est celui où un mot est toujours accompagné par un autre mot, le «mot ciblé», ce

[5] Principe de compositionnalité: «Le sens d'un énoncé dépend du sens de ses parties et de la façon dont celles-ci sont combinées dans l'énoncé» (http://www.sir.blois.univ-tours.fr/~antoine/enseignement/tal/crs_semantique.pdf)

qui donne la possibilité de déterminer le sens, en fonction de la succession, la distance et la relation syntaxique. Le mot *co-occurrent* le plus fréquent est considéré la *collocation* la plus pertinente.

Les nouveaux systèmes de désambiguïsation du sens des mots emploient des techniques d'apprentissage inductif (Tătar, 2003) à partir des corpus de données désambiguïsées et annotées. Ces techniques permettent d'extraire, pour un mot donné, tous les contextes annotés avec le sens correct, en utilisant les *attributs contextuels*. Ces attributs peuvent être: des *collocations* – liées par un rapport de proximité syntaxique mais aussi par des positions bien déterminées à gauche ou à droite (en mesurant la distance entre deux éléments) ou des *co-apparitions* (*co-occurrences*) liées par un rapport de relation syntaxique mais avec une distance relative entre les mots. Dans ce cas on peut choisir comme attributs les plus fréquents *n* co-apparitions (co-occurrences) d'un mot.

On peut définir des «multi-ensembles» de mots, en calculant la probabilité de l'apparition d'un mot dans un certain contexte. Par exemple, pour le mot «mémoire», conformément au dictionnaire on peut y avoir plusieurs sens. Alors, on peut définir un ensemble avec les éléments: *écrit*, *sommaire*, *idée,* etc., qui donne le sens *s1*, et un autre ensemble d'éléments: *ordinateur*, *logiciel*, *données,* etc. qui peut donner un autre sens *s2*. Si le mot «mémoire» apparaît dans le même contexte que le mot ordinateur, on peut affirmer qu'il y a une grande probabilité que le sens du mot soit *s2*. La constitution des multi-ensembles peut être automatisée en employant les dictionnaires électroniques disponibles.

En ce qui concerne l'ambiguïté syntaxique, les situations les plus fréquentes sont celles qui se réfèrent aux subordonnées relatives et aux syntagmes prépositionnels. On peut déterminer, à l'aide des calculs probabilistiques, si un substantif ayant une signification établie est sujet ou complément pour un certain verbe. De plus, avec ces méthodes on peut déterminer la structure des syntagmes nominaux qui contiennent plusieurs syntagmes prépositionnels. Par exemple, pour le syntagme «machine à laver les pommes de terre à tambour» la structure attribuée sera [[machine [à laver les pommes de terre] [à tambour]]] et non l'autre variante [machine [à laver les pommes de terre [à tambour]].

Dans les analyses des propositions simples ces méthodes et modèles mathématiques ont une grande efficacité. Par contre, pour les propositions plus complexes, – les phrases, – d'autres problèmes spécifiques au langage naturel interviennent tels, par exemple, *l'anaphore*. Les problèmes de la résolution de l'anaphore ont une grande importance notamment dans le domaine de la traduction automatique, de la recherche/l'extraction des informations (pour l'indexation des documents), du résumé automatique, etc. Les algorithmes (Carbonell, Brown, 1988 ; Lappin, 1994 ; Mitkov, 1994 ; Mitkov, 1995) sont basés sur des méthodes traditionnelles (syntaxiques et sémantiques) qui sont combinées avec des méthodes plus récentes de la pragmatique et de la théorie du discours.

Processus de catégorisation et extraction des informations

L'information contenue dans un texte/document doit être présentée sous une forme intelligible pour que le texte/document soit facilement et correctement perçu. Dans l'exploitation de celui-ci il y a plusieurs étapes: la création, le traitement, la distribution et l'utilisation. Parmi les applications informatiques qui peuvent être utilisées dans le cas des documents, la *catégorisation* et l'*extraction* des informations sont très importantes et surtout, d'une grande actualité dans les recherches textuelles.

Le processus de catégorisation des textes

La catégorisation est l'un des processus fondamentaux de l'activité cognitive qui se manifeste dans l'activité du langage, notamment dans celle du raisonnement (Sarrasin, 2005). D'après Moechler et Auchlin (2005:51) la catégorisation signifie, en fait, l'assignation d'un objet (ou processus) à une classe d'objets (processus) ou à un ensemble de classes. Pour être catégorisés les objets ou processus doivent contenir toutes les propriétés de la classe (condition nécessaire et suffisante).

Les catégories peuvent être structurées selon une hiérarchie avec plusieurs niveaux: des niveaux généraux (donc plus abstraits) et des niveaux subordonnés (plus spécifiques). Les études ont montré que, parmi ces niveaux, le niveau de base est le plus utilisé parce qu'il contient le plus grand nombre d'attributs[6]. Il y a des recherches très sérieuses dans le domaine des sciences cognitives qui étudient la modalité par laquelle le cerveau humain utilise les processus de catégorisation pour les concepts dans divers contextes et la manière dont ces processus pourraient être automatisés (appliqués aux ordinateurs) (Sarrasin, 2005; Delacour, 2001).

À partir de cela, on peut considérer la méthode comme un groupement des informations dans les différentes catégories, en fonction du degré de ressemblance ou en fonction du type d'association.

Dans le domaine de la technologie de l'information, la catégorisation est définie comme le «processus d'attribution des objets à une ou plusieurs classes ou catégories», avec le plus grand taux de réussite possible.

Le modèle vectoriel est l'un des modèles couramment utilisé. Un document peut être considéré comme un vecteur d'attributs, et le processus d'interrogation est à son tour un vecteur d'attributs (Manning, Schütze, 1999; Jackson, Moulinier, 2002; Tătar, 2003). Pour le cas du texte/document on peut définir un *modèle de classe*[7] (avec les paramètres de la classe) et aussi une *procédure* (qui fait par exemple la sélection pour la classe). L'algorithme de recherche des informations utilise une méthode de comparaison: il compare son vecteur d'attributs avec le vecteur du document et désigne une liste de documents «appropriés». Les calculs

[6] «Ce niveau de base se manifesterait en raison du fait qu'il maximise le potentiel informatif des concepts» (Sarrasin, 2005).

[7] Modèle de représentation des données le plus convaincant.

sont influencés par la fréquence de l'apparition des termes ou, par contre, par le nombre réduit d'apparitions.

Au début, les méthodes de catégorisation des textes étaient conçues sur la base d'un ensemble de règles élaborées manuellement (*knowledge engineering*); les règles étaient sous la forme (Jackson, Moulinier, 2002, Tătar, 2003):

if *<formule Booléen sous la forme (AND, OR, NOT) >* ***then*** *<catégorie>*

Avec le développement d'Internet, dans les techniques du processus de catégorisation ont été introduites des techniques d'apprentissage automatique (*machine learning)*. À partir des différentes catégories de documents on peut créer, par ces méthodes, un classificateur propre, par un processus inductif d'observation des caractéristiques de l'ensemble des documents (qui sont préalablement classifiés). La classification automatique des textes consiste à attribuer une catégorie à chaque texte. Après la méthode dont les classes sont générées la classification peut être : supervisé (les groupes des documents ou classes sont calculé automatiquement par un expert), ou une classification non supervisé (les groupes des documents sont calculé automatiquement par la machine).

La catégorisation s'applique aux situations de classification des textes (documents) en fonction du sujet, à l'identification des informations qui appartiennent à certains auteurs, à la désambiguïsation du sens du mot, à la détection des courriels non sollicité (*spams*), à l'identification de la langue d'un texte écrit, au processus de recherche des informations, etc.

L'extraction des informations

À la différence des techniques de recherche de documents, le processus d'extraction des informations se réfère à l'extraction des *données essentielles* contenues dans les documents et les textes divers. Par ces méthodes on peut créer des classifications complexes des documents qui contiennent un certain sujet, on peut créer des ensembles de documents qui contiennent certains concepts, on peut générer des bases de métadonnées, on peut extraire certaines informations des pages web spécialisées ou moins spécialisées.

Ces méthodes sont basées sur des techniques d'analyse des mots associée avec des contraintes linguistiques, pour l'identification des mots clés, mais il y a aussi des systèmes qui utilisent de graphes conceptuels. Ces systèmes doivent avoir une grande rapidité de traitement des mots si on veut appliquer cette technologie aux grands ensembles de textes semi ou non structurés. Les modèles utilisés sont ceux des automates à l'état fini, des grammaires indépendantes du contexte, de la reconnaissance des modèles et de la modélisation statistique (Jackson, Moulinier, 2002).

Les méthodes ont certes évolué avec le développement des documents électroniques disponibles en ligne et la constitution des grands corpus (surtout les corpus de documents scientifiques), et récemment des corpus de textes commerciaux (moins structurés).

Les données linguistiques sont organisées en bases de données spécifiques qui font appel aux bases de connaissances, bases de données annotées (représentées sous forme d'automates). Les automates sont décrits par des algorithmes qui peuvent servir à construire des systèmes d'interrogation ou systèmes de dialogue.

Ces méthodes contiennent les étapes suivantes (Cole, 1996): 1. *identification des artéfacts des textes*, des mots clés, noms propres, dates, lieu, etc., par une méthode d'écrêtage (*text skimming*) – (parcourir très rapidement un texte pour l'identification des idées principales); 2. utilisation *des contraintes linguistiques;* 3. *utilisation des bases de connaissance*s pour l'identification des contenus.

Soit, par exemple (Tătar, 2003), un texte spécialisé (domaine des affaires), dont on peut extraire des informations sur le partenariat, par les étapes suivantes: identification du nom des compagnies, identification des mots clés qui décrivent les affaires (en utilisant des bases de connaissances – et sachant, par exemple, que les fusions impliquent deux partenaires). On peut extraire des informations sur chiffres les d'affaires, relations, partenariats, etc. structurées sous diverses formes.

L'identification de certaines expressions du texte, pour diverses langues, est basée sur des méthodes de filtrage par modèle (motif) (*pattern matching*), en utilisant des *expressions régulières*[8].

Les analyses grammaticales peuvent être raffinées par l'introduction des lexiques, thésaurus, bases de données organisées en catégories (pour une analyse sémantique primaire). Les techniques nouvelles utilisent l'apprentissage automatique à l'aide des algorithmes qui doivent découvrir la structure des données. Ainsi, de grands corpus de textes sont analysés, segmentés et annotés. À partir de cela sont générées des règles de comportement pour les algorithmes.

Par exemple (Jackson, Moulinier, 2002:109), la proposition:

'The parliament was bombed by Carlos.'

peut être annotée:

"The <TARGET>parliament</TARGET> was
<ACTION>bombed</ACTION> by <PERP>Carlos</PERP>"

Avec ces étiquettes un logiciel pourrait créer un modèle:

"NOUN was PASSIVE-VERB by NOUNGROUP"

Comme méthodes alternatives on utilise les modèles statistiques qui emploient des processus d'analyse statistique et probabilistique.

Le processus d'extraction des informations peut être utilisé aussi pour le traitement des ontologies[9] (identification des diverses relations des données et extraction des informations).

---

[8] Une *expression rationnelle* ou *expression régulière* est «une chaîne de caractères que l'on appelle parfois un motif et qui décrit un ensemble de chaînes de caractères selon une syntaxe précise». (http://fr.wikipedia.org/wiki/)

[9] Ontologie (dans le domaine de l'informatique) est «un système de représentation des connaissances» (http://fr.wikipedia.org/wiki)

Un logiciel très connu est *Tipster* – un système d'extraction des informations de la presse britannique et japonaise. Il y a aussi *Crystal, Palka*, *Hasten* (Muslea, 1999), des logiciels automatisée, qui utilisent l'extraction des informations à partir des algorithmes. *MindNet* est un logiciel de représentation des connaissances qui, basé sur un analyseur grammatical et sur des dictionnaires, encyclopédies ou corpus de textes, peut construire des réseaux sémantiques. Le processus, entièrement automatisé fait une analyse de chaque phrase et établit des graphiques sémantiques et d'autres calculs de probabilité.

Conclusions

L'article présente quelques méthodes et techniques employées dans le processus de catégorisation et d'extraction des informations. Ces modèles, considérés techniques de «modélisation des information» sont basés, en général, sur des théories linguistiques d'analyse du langage (syntaxiques et sémantiques) mais aussi sur des méthodes statistiques, probabilistes. La description et la mise en œuvre de ces techniques ne se limitent pas au domaine de l'informatique. Ces processus sont essentiels pour le domaine de la traduction automatique, mais également pour les systèmes d'aide à la traduction. Les techniques contiennent des analyses grammaticales, des recherches et des modalités de classification des éléments, afin de créer des *modèles d'analyse translinguistique* et, récemment, des analyses statistiques de pointe pour la sélection et la réorganisation des mots conformément à l'ordre des mots de la langue cible.

La formation et l'évolution du langage sont toujours associées par des liens étroits et complexes au processus de raisonnement, au processus de *construction* et/ou d'*élection* d'une variante plus pertinente (optimale) parmi les infinités des alternatives. En conséquence, les processus de *catégorisation* et d'*extraction* des informations, considérés comme processus fondamentaux de l'activité cognitive, sont essentiels.

L'analyse de ces travaux rend possibles les études complexes de l'activité cérébrale sur les divers modes de représentations des données et, notamment, des connaissances.

BIBLIOGRAPHIE

BROWN, F.P., P.V. DESOUZA, R.L. MERCER, V.J. DELLA PIETRA, J.C. LAI (1992) Class-Based n-gram Models of Natural Language. In *Computational Linguistics*, vol 18(4), 1992, pp. 467-480, http://acl.ldc.upenn.edu/J/J92/J92-4003.pdf

CARBONELL, J.G., R.D. BROWN (1988), Anaphora Resolution: A Multi-strategy Approach. In *International Conference on Computational Linguistic. Proceeding of the 12th Conference on Computational Linguistic*, vol. 1, pp. 96-101. http://www.aclweb.org/anthology-new/C/C88/C88-1021.pdf

CHOMSKY, N. (1996) *Cunoaşterea limbii.* Bucureşti: Editura Ştiinţifică

CHOMSKY, N., G. MILLER (1969) *L'analyse formelle des langues naturelles*. Paris: Éditions Mouton

COLE, RONALD, A., (ed.) (1996) *Survey of the State of the Art in Human Language Technology*. Cambridge: University Press http://cslu.cse.ogi.edu/HLTsurvey

CRISTEA, D. (2002) *Formalisme şi instrumente de descriere şi prelucrare ale limbajului natural*. Iaşi: Editura Universităţii Al. I. Cuza

DELACOUR, J. (2001) *Introducere în neuroştiinţele cognitive*. Bucure□ti: Editura Polirom

ECO, U. (2007) *Limitele interpretării*. Ia□i: Editura Polirom

GALES, M., S. YOUNG (2008) The Application of Hidden Markov Models in Speech Recognition. In *Foundations and Trends® in Signal Processing*: Vol. 1(3), 2007, pp 195-304

GOLU, M. (1975) *Principii de psihologie cibernetică*. Bucureşti: Editura Ştiinţifică şi Enciclopedică

JACKSON, P, I. MOULINIER (2002) *Natural Language Processing for on line Applications: Text Retrieval, Extraction and Categorization*. Amsterdam: John Benjamins Publications

JURAFSKY, D., J.H. MARTIN (2008) *Speech and Language Processing*. Prentice Hall

LAPPIN, S., H.J., LEASS (1994), An Algorithm for Pronominal Anaphora Resolution. In *Computational Linguistic*, vol. 20(4), 1994, http://acl.ldc.upenn.edu /J/J94/J94-4002.pdf

LAPPIN, S., (ed.) (1996) *The Handbook of Contemporary Semantic Theory*. Oxford, UK, Malden, Massachuttes, USA: Blackwell Publishers

MANNING, C., H., SCHÜTZE (1999) *Foundation of Statistical Natural Language Processing*. Massachusetts: MIT Press

MIHALCEA, R., T. PEDERSEN (2005) Advances in Word Sense Disambiguation. Tutorial at *Twentieth National Conference on Artificial Intelligence (AAAI-05)*, July, 9-13, 2005, Pittsburg, Pennsylvania, http://www.d.umn.edu/~tpederse/ WSDTutorial.html

MITKOV, R., S-K., CHOI (1995) Anaphora Resolution in Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*, Leuven, Belgium, http://www.mt-archive.info/TMI-1995-Mitkov.pdf

MITKOV, R. (1994) An integrated model for anaphora resolution. In *Proceedings of the 15th International Conference on Computational Linguistics COLING'94*, Kyoto, Japan, 1994, http://aclweb.org/anthology-new/C/C94/C94-2191.pdf

MOESCHLER, J. A., AUCHLIN (2005) *Introducere în lingvistica contemporană*. Cluj-Napoca: Editura Echinox

MUSLEA, I., (1999) Extraction patterns for information extraction tasks: A survey. In Papers from the AAAI Workshop on *Machine Learning for Information Extraction*, http://www.isi.edu/~muslea/PS/ml4ie-aaai99.pdf.

PERRAUD, F., E. MORIN, C. VIARD-GAUDIN, P.M., LALLICAN (2003) Apport d'une modèle de language statistique pour la reconnaissance de l'écriture manuscrite en ligne. In *Actes de Traitement Automatique de Langues Naturelles, TALN 2003*, vol 1, http://www.sciences.univ-nantes.fr/info/recherche/taln2003/articles /perraud.pdf

SARRASIN, N. (2005) Tirer profit du modèle cognitif humain dans les recherches en intelligence artificielle. In *Automates Intelligents* n° 60, août 2005

SAUSSURE, F., DE (1998) *Curs de lingvistică generală*. Bucureşti: Editura Polirom

TĂTAR, D. (2003) *Inteligență artificială. Aplicații în prelucrarea limbajului natural*. Cluj-Napoca: Editura Albastră

VASILIU, E. (1970) *Elemente de teorie semantică a limbilor naturale*. Bucureşti: Editura Academiei Republicii Socialiste România

VASILIU, E., S. GOLOPENȚIA-ERETESCU (1969) *Sintaxa transformațională a limbii române*. Bucureşti

VASILIU, E., S. STATI, S. GOLOPENȚIA (2003) *Introducere în teoria lingvistică*, http://www.unibuc.ro/eBooks/filologie/dominte/10-3.htm

VRACIU, A. (1980) *Lingvistică generală şi comparată*. Bucureşti: Editura Didactică şi Pedagogică

WINOGRAD, T. (1987) A Language/Action Perspective on the Design of Cooperative Work. In *Human-Computer Interaction* 3:1 (1987-88), 3-30, http://hci.stanford.edu/ ~winograd/papers/ language-action.html

Manuela MIHĂESCU is Assistant Lecturer at the Applied Modern Language Department of the Faculty of Letters, Babeş-Bolyai University of Cluj-Napoca, Romania. Currently she is pursuing her doctoral studies in linguistics and is also involved in European research projects on Romanian language processing, syntagma processing. Her research interests are also in communication and multimedia processing.

Sanda CHERATA is Assistant Lecturer at the Modern Applied Languages Department of the Faculty of Letters, Babeş-Bolyai University of Cluj-Napoca, Romania. Her research includes natural language processing, computational linguistics, language engineering, and terminology. She is currently involved in projects regarding Romanian language processing and terminological databases.