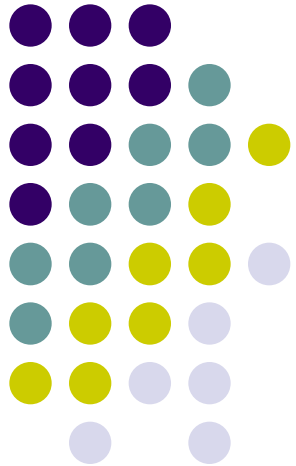
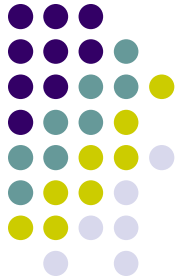
SCI6134
Outils linguistiques et gestion documentaire



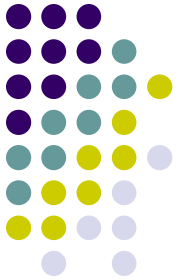
Le traitement automatique de la langue et sa place dans la chaîne documentaire

Plan du cours



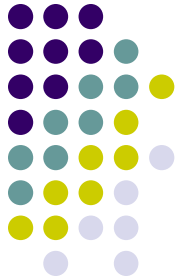
- Le TAL (traitement automatique de la langue)
- La place du TAL dans la chaîne documentaire

Le TAL



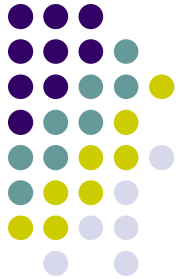
- Objectif du TAL
- Méthodologie
- Justification (pourquoi faire du TAL?)
- Étapes de traitement
- Petit historique

Le TAL : objectif



- Objectif : extraire automatiquement du texte en langue (naturelle) des informations pour effectuer un traitement au document, ex. :
  - traduction automatique
  - correction automatique
  - indexation automatique
  - acheminement automatique du courriel
- Langue écrite vs. langue parlée
- Langue (naturelle) vs langage artificiel

Le TAL : méthodologie

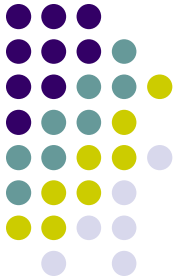


- Méthodologie habituelle
  - chaînes de caractères → représentation sémantique ou autre
- Représentation finale : permet le traitement « intelligent »
- Exemple (traduction automatique)
  - *The girl misses the boy* → Le garçon manque à la fille
  - comment? relier chaque mot à sa traduction, là où c'est possible; ajuster les structures syntaxiques

Le TAL : méthodologie *(suite)*

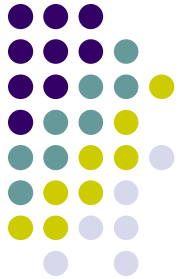
Problème de traduction	**Source**	**Cible**
Mots sans traduction	*lap*	?
	chez	?
Réalités culturelles sans équivalent	*latchkey children*	?
Découpages différents de la réalité selon les langues	*river*	rivière/fleuve
Expressions idiomatiques	sans tambours ni trompettes	*without any fuss*, *unobstrusively*
Phraséologie différente	Thomas aime nager	*Thomas schwimt gerne*
	I drove to the airport	J'ai pris ma voiture pour me rendre à l'aéroport. Je suis allé à l'aéroport.

Le TAL : justification



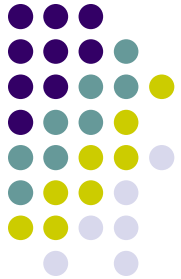
- Théorique : vérifier des théories linguistiques
- Pratique : développement de systèmes utiles (ex., Bouillon et al., pp. 7-8)
  - traduction automatique
  - indexation automatique
  - recherche documentaire
  - résumé automatique
  - systèmes de dialogue homme-machine
  - systèmes de questions-réponses
  - correction automatique
  - génération de textes

Le TAL : étapes de traitement classiques



1. Identification des mots, ponctuation, nombres; phrases; paragraphes. Exemples :
   - homme-grenouille
   - M. le président Louis D. Lamontagne
   - Qu'a-t-il dit ? qu'il reviendrait. À 8h au plus tard.
2. Analyse des mots (analyse morphologique)
   - ex. hommes; reviendrait; homme-grenouille; a-t-il
3. Analyse syntaxique : reconnaître la structure de la phrase, les liens entre les mots de la phrase
   - David a battu Goliath; Goliath a été battu par David
   - Pierre est allé étudier la linguistique à Paris et Marie la littérature américaine à Boston.

Le TAL : étapes de traitement classiques *(suite)*

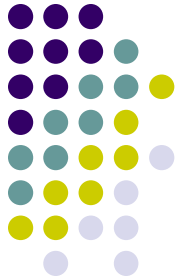


4. Analyse sémantique : identifier le sens de la phrase, notamment pour pouvoir la traduire
   - David a battu Goliath vs. Goliath a battu David
   - Tous les étudiants de la classe parlent deux langues.

     vs.

     Deux langues sont parlées par tous les étudiants de la classe.
   - Le chien et le chat sont des animaux de compagnie mais le félin est le plus indépendant des deux.

Le TAL : étapes de traitement classiques *(suite)*



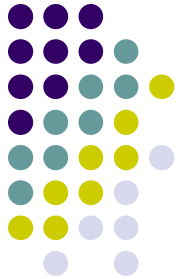
5. Pragmatique
   - La mère a grondé sa fille parce qu'elle a traversé la rue toute seule.

     vs

     La mère a grondé sa fille parce qu'elle avait perdu patience.
   - Je vais distribuer des canapés aux avocats.

- Aussi (préalable) : normalisation des formats de fichier; interprétation des balises de formatage autour du texte réel; etc.

Le TAL : étapes de traitement classiques *(suite)*



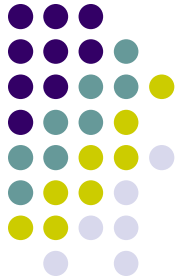
- Méthodes superficielles
  - repérage de syntagmes (groupes) nominaux
  - *chunking* (repérage de groupes de mots importants – surtout les groupes nominaux -, les verbes conjugués, quelques relations)
  - identification d'entités nommées (voir plus tard)
  - statistiques lexicales

Le TAL : petit historique

- Histoire du TAL très lié à celle de la traduction automatique (voir Hutchins, 1986; Somers, 2011)
- Débuts : ~1950
- Effort initial : traduction automatique – russe-anglais
- Rapport ALPAC (1966) :

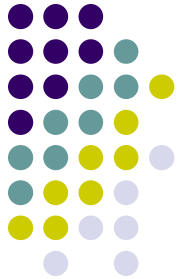
  « les recherches dans leur état actuel ne sont pas rentables pour l'État Américain. Les subsides pour la TA sont donc coupés aux USA du jour au lendemain. » (Bouillon et al., 1998, p. 10)
- Recherche fondamentale sur le TAL (approches symboliques)
  - influence de Chomsky et de la linguistique générative, i.e. formelle
- Années 1970 : quelques systèmes simplistes mais étonnants nourrissent l'intérêt pour la recherche

Le TAL : petit historique *(suite)*



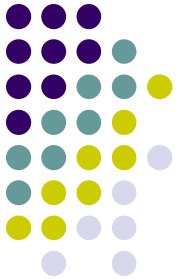
- Travaux durant les années 1980 : fortement influencés par ceux en intelligence artificielle
  - représentation des connaissances, raisonnement, etc.
- ~ 1990 : approches probabilistes
- Depuis mi-1990
  - importance croissante à cause d'Internet (moteurs de recherche de plus en plus sophistiqués)
  - applications de gestion documentaire gagnent du terrain sur la traduction automatique
- Aujourd'hui : apprentissage automatique

Plan du cours *(rappel)*



- Le TAL (traitement automatique de la langue)
- La place du TAL dans la chaîne documentaire

La place du TAL dans la chaîne documentaire



- Introduction
- Importance du TAL dans la gestion documentaire
- Liens entre le TAL, l'informatique et les sciences de l'information

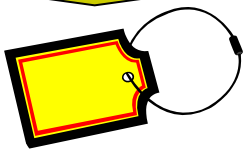
TAL et chaîne documentaire : introduction

population de documents

sélection

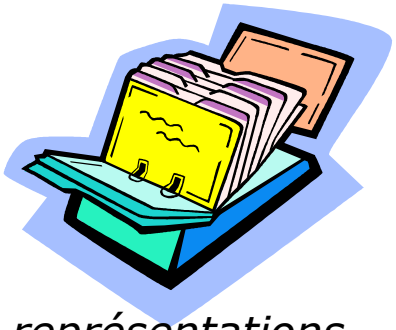
collection de documents

description



notices bibliographiques

analyse documentaire



représentations du contenu

diffusion

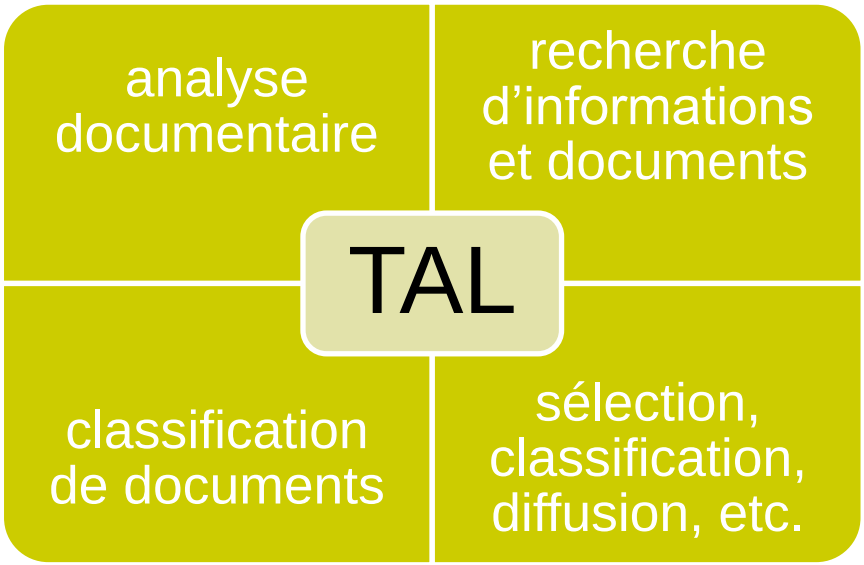
utilisateurs

stockage

accès

TAL et chaîne documentaire : introduction *(suite)*

- Transformations dans le traitement documentaire
  - causes : micro-informatique et collections de textes numériques
- Impact du TAL : là où le texte est en jeu
- Occasion en or pour la linguistique informatique
- Défi : modéliser toutes les dimensions de la gestion documentaire



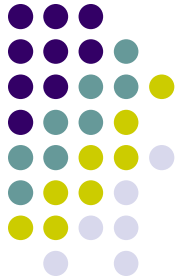
TAL et chaîne documentaire : importance du TAL

- Perçue très tôt
  - Masterman et al. (1958); Sparck Jones (1967)
  - particulièrement : repérage de documents (*information retrieval*)
- Croissante

Publications	• Ambroziak et Woods, 1998; Strzalkowski 1999; Voorhees 1999; Perez-Carballo et Strzalkowski, 2000; Oard et al., 2001; Todirasçu et Rousselot, 2001; Ruch, 2003; Lalich-Bodin et Maret, 2005; Normier, 2007; Dong et al., 2008; Formica et al., 2016
Congrès	• NLP4DL (2009); NLP4ITA (2013); Workshop on Health Text Mining and Information Analysis (2014)

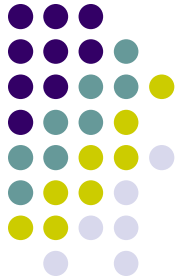
- Liens importants avec le Web sémantique
- Enjeux très importants avec la valeur économique de l'information numérique

TAL et chaîne documentaire : importance du TAL *(suite)*



- Terminologie (ou courants de recherche)
  - indexation automatique, semi-automatique ou assistée par ordinateur
  - résumé/condensation automatiques, semi-automatiques ou assistés par ordinateur
  - classification/catégorisation automatique, *clustering*
  - recherche/repérage d'information
  - extraction/repérage/filtrage de termes
  - forage textuel, fouille de textes
  - extraction d'information
  - systèmes de questions-réponses
  - construction automatique de thésaurus
  - annotations pour le Web sémantique
  - repérage d'images et d'autres contenus non textuels
  - etc.

TAL et chaîne documentaire : TAL, informatique et SI



- Applications initiales issues de l'informatique avec peu ou pas de contribution des sciences de l'information
  - *automatic summarization*, etc.
- Applications nouvelles (ex. fouille de textes, constitution automatique de thésaurus, analyses de contenu, Web sémantique, etc.)
- Conditions nécessaires pour atteindre la synergie à prévoir
  - pas encore complètement réunies
  - se concrétisent graduellement
  - incluent l'implication des sciences de l'information

Références

- Ambroziak, Jacek; Woods, William A. 1998. Natural Language Technology in Precision Content Retrieval. In : *Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA 98)*, August 18-21, 1998, Moncton, New Brunswick, Canada.
- Bernardi, Raffaella; Chambers, Sally; Gottfried, Björn; Segond, Frédérique ; Zaihrayeu, Ilya (réds.). 2009. In : *Advanced Language Technologies for Digital Libraries - International Workshops on NLP4DL 2009*, Viareggio, Italy, June 15, 2009 and AT4DL 2009, Trento, Italy, September 8, 2009. Lecture Notes in Computer Science 6699 Springer 2011.
- Bouillon, Pierrette et al. 1998. *Traitement automatique des langues naturelles.* Paris ; Louvain-la-Neuve : Duculot.
- Dong, Hai; Hussain, Farookh; Chang, Elizabeth. 2008. A survey in semantic search technologies. *2nd IEEE International Conference on Digital Ecosystems and Technologies, IEEE-DEST 2008*. 403 - 408. doi : 10.1109/DEST.2008.4635202.

Références

- Formica, Anna; Missikoff, Michele; Pourabbas, Elaheh ; Taglino, Francesco. A Bayesian Approach for Weighted Ontologies and Semantic Search. In : *IC3K 2016 Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pp 171-178.
- Hutchins, W.J. 1986. *Machine Translation: Past, Present, Future*. Chichester : Ellis Horwood.
- Lallich-Bodin, G.; Maret D. 2005. *Recherche d'information et traitement de la langue : fondements linguistiques et applications*. Lyon : Les Presses de l'enssib.
- Masterman, M.; Needham, R.M.; Sparck Jones, K. 1959. The analogy between mechanical translation and library retrieval. In : *Proceedings of the International Conference on Scientific Information* (1958), National Academy of Sciences – National Research Council, Washington, D.C., Vol. 2, pp. 917-935.

Références *(suite)*

- Moens, Marie-Francine. 2000. Automatic indexing and abstracting of document texts. Boston : Kluwer Academic Publishers.
- Nazarenko, Adeline. Le point sur l'état actuel des connaissances en traitement automatique des langues (TAL). In Sabah (réd.) Compréhension des langues et interaction, Paris : Hermès, 2006, pp. 31-70.
- Normier, Bernard. 2007. L'apport des technologies linguistiques au traitement et à la valorisation de l'information textuelle. Paris : ADBS Éditions.
- Oard, Douglas W. et al. 2001. Multilingual Information Retrieval. In : Hovy, E.; Ide, N.; Frederking, R.; Marian, J.; Zampolli, A. (réds.), Multilingual Information Management: Current Levels and Future Abilities, pp. 223-256.
- Perez-Carballo, J.; Strzalkowski, T. 2000. Natural language information retrieval: progress report. Information Processing & Management 36(1):155-78.
- Sparck Jones, Karen. 1967. Current work on automatic classification for information retrieval. T.A. Informations, 2:92-96.

Références *(suite)*

- Strzalkowski, T. (réd.). 1999. *Natural Language Information Retrieval*. Dordrecht : Kluwer Academic Publishers.
- Ruch, Patrick. 2003. *Applying Natural Language Processing to Information Retrieval in Clinical Records and Biomedical Texts*. Thèse de doctorat. Genève : Imprimerie des Hôpitaux Universitaires de Genève.
- Somers, Harold. 2011. Machine Translation: History, Development, and Limitations. In : Malmkjær, Kirsten; Windle, Kevin (réds) *The Oxford Handbook of Translation Studies* (ch. 28), Oxford University Press.
- Todirasçu, Amalia; Rousselot, François. 2001. Ontologies for Information Retrieval. In : *TALN 2001*, Tours, 2-5 juillet 2001. [http://tln.li.univ-tours.fr/Tln_Colloques/TALN2001-RECITAL2001/Actes/tome1_PDF/partie2_p30_322/art28_p303_312.pdf]
- Voorhees, E.M. 1999. Natural language processing and information retrieval. In : Pazienza, M.R. (réd.) *Information extraction. Towards scalable, adaptable systems*, Berlin: Springer-Verlag, pp. 32-48.