TAL et Recherche d'Information
profilage linguistique des requêtes

Ludovic TANGUY (CLLE-ERSS Axe TAL)
et Josiane Mothe (IRIT Equipe SIG)

Vue d'ensemble

- Collaboration entre RI et TAL
  - Des affinités évidentes…
  - …mais un mariage malheureux
  - Une relation à réévaluer et à réinventer
- Une expérience : le profilage linguistique des requêtes
  - Pour prédire la difficulté
  - Pour adapter les traitements aux données
  - Application aux traitements morphologiques

Plan

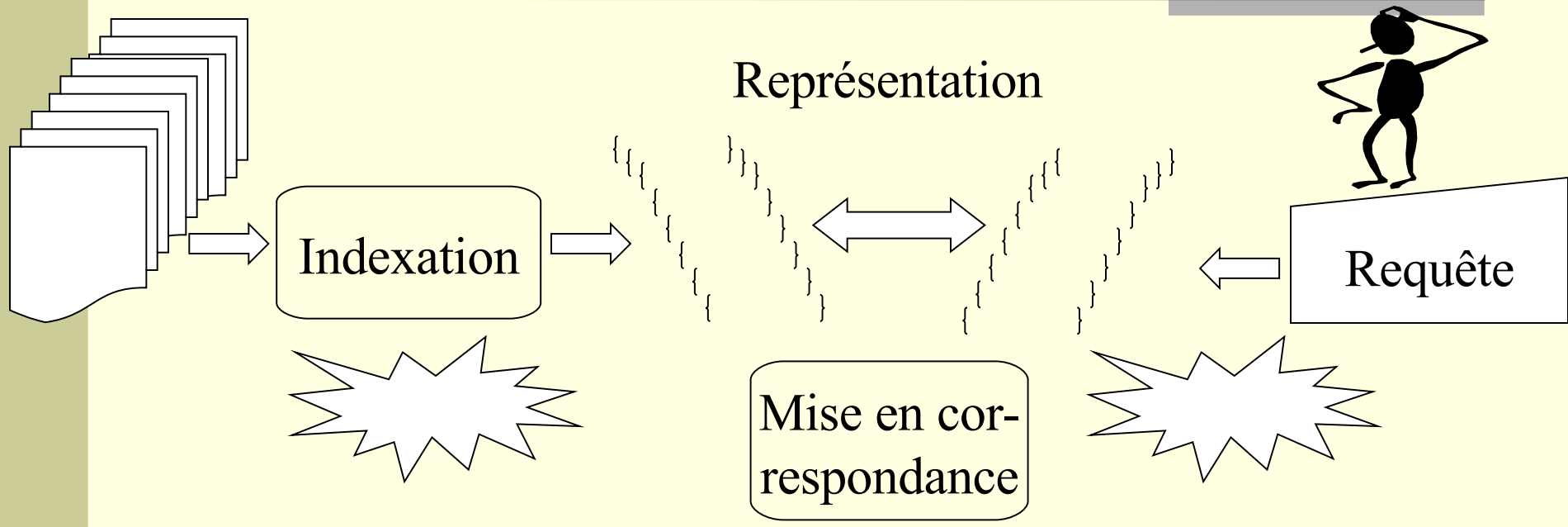
- Généralités sur la RI
  - Rapports entre TAL et RI
  - Campagnes d'évaluation (TREC/CLEF)
  - La plate-forme RFIEC de l'IRIT
- Phase d'observation
  - Exploitation des archives des campagnes
  - Examen local des mécanismes d'une chaîne de RI
- Phase d'action
  - Profilage des requêtes
  - Fusion de systèmes
  - Vers un système adaptatif

Quelques généralités sur la RI et le TAL

La RI en quelques mots

- Étant donné :
  - Une collection de documents
  - Un besoin exprimé (requête)
- Retrouver les documents correspondant à ce besoin
  - Le plus précisément et le plus exhaustivement possible

Schéma général



Caractéristiques du modèle dominant

- Modèle vectoriel (VSM)
  - Représentation des documents et des requêtes par « sac de mots »
  - Pondérés par leur fréquence relative
- Axes d'amélioration
  - Mesures pour la mise en correspondance et le classement (paramétrages)
  - Boucles de rétroaction (blind feedback)
  - Approches statistiques opaques des proximités lexicales (LSI)

Utilité du TAL pour la RI

- Extraction des unités pertinentes (docs+requêtes)
  - Segmentation (mots / syntagmes)
  - Analyseurs automatiques (segmenteurs, chunkers, taggers, parsers)
  - Normalisation des formes (typographie, morphologie, syntaxe)
  - Passage aux concepts (désambiguïsation)
- Expansion de requêtes
  - Ajout de termes à ceux de la requête initiale, sur critères morphologiques ou sémantiques
  - Ressources lexicales (BD morphologiques, réseaux sémantiques, etc.)
- Mesures de similarité
  - Proximité lexicale, modèles de langue

Utilité de la RI pour le TAL

- La RI comme valorisation des productions du TAL
  - Lexiques, réseaux, analyseurs, etc.
  - Invocation de la RI comme terrain applicatif et utilité sociale
- Les campagnes d'évaluation de la RI comme terrain de jeu pour le TAL (TREC, CLEF, etc.)
  - Collections de données disponibles
  - Evaluation des techniques suivant des critères reconnus

Bilan mitigé (pour le moins)

- Ellen Voorhes (1999) :
  - Currently, the most succesful general purpose retrieval systems are statistical methods that treat text as little more than a bag of words. However, attempts to improve retrieval performance through more sophisticated linguistic processing have been largely unsuccessful.
- Claude de Loupy (2001) :
  - Les expériences publiées dans la littérature ne font pas apparaître clairement que les systèmes utilisant des connaissances linguistiques obtiennent de meilleures performances.
- Thorsten Brants (2003) :
  - Many NLP techniques have been used in IR. The results are not encouraging. Simple methods (stopwording, Porter-style stemming, etc.) usually yield significant improvements, while higher-level processing (chunking, parsing, WSD) only yield very small improvements or even a decrease in accuracy. At the same time, these methods increase the processing and storage costs dramatically.

Des problèmes de couple ?

- Différentes constatations et explications :
  - Peu de publications communes des deux communautés
  - Les techniques de TAL ne sont pas suffisamment au point
  - Les tâches génériques de RI sont difficiles pour le TAL
  - La RI utilise des techniques de TAL "prêtes à l'emploi", puis applique ses propres méthodes
- Qu'est-ce qu'un bon système de RI ?
  - Quelques doutes sur la notion d'efficacité en ce qui concerne les techniques de TAL

Les campagnes d'évaluation

- Communauté scientifique mobilisée très tôt pour la mise en commun de ressources
- Text REtrieval Conference (TREC), depuis 1992
  - Langue anglaise essentiellement
- Cross-Language Evaluation Forum (CLEF), depuis 2000
  - Autres langues européennes et recherche d'information trans-langue
- NII-Nacsis Test Collections for Information Retrieval (NTCIR), 1998
  - Langues asiatiques
- Passages obligés pour tout système de RI

Évaluation « à la TREC »

- Plusieurs tâches
  - Adhoc, WEB, Hard, QA, Terabyte, SPAM, Novelty, etc.
  - Collections de documents (presse, Web, mails, etc.)
  - Collection de requêtes (« topics »), 50 par tâche par an
    - Texte de la requête (plus ou moins structuré)
    - Liste des documents « pertinents »
- Chaque campagne (annuelle)
  - Définition de la collection
  - Distribution des requêtes
  - Exécution des recherches par le système évalué (différents paramètres possibles, ou « runs »)
  - Retour des résultats et évaluation du « run »

Exemple de requête (adhoc, topic 35, 2003)

Title: NATO, Poland, Czech Republic, Hungary

Descriptive: Accession of new NATO members: Poland, Czech Republic, Hungary, in 1999.

Narrative: Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant. Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

Notion de pertinence

- Décision humaine binaire pour un couple document/requête
- « Pooling method » : les juges ne se basent que sur un sous-ensemble de documents
  - Uniquement ceux ramenés par au moins un système
- Pas de méthodologie claire ni de classement pour les documents pertinents

Une évaluation « globale »

- Plusieurs scores globaux (moyenne) pour chaque « run »
  - Classement des systèmes sur cette base
  - Multiplicité des mesures
- Les scores pour chaque requête sont disponibles, mais non utilisés
  - Uniquement une moyenne sur les requêtes
- Au final, très peu de différences entre les « bons » systèmes
  - Même si des variations importantes peuvent exister d'une requête à l'autre

Conséquences

- Peu de visibilité des difficultés rencontrées par les systèmes
- Pas de prise en compte de l'efficacité par rapport à une requête « difficile »
- Pas de retour précis sur les techniques utilisées
  - Peu d'information publiée sur les caractéristiques des systèmes
  - Les techniques de TAL peuvent avoir des effets non mesurables suivant ces critères globaux

Études locales souhaitables

- Observer le comportement d'un système par rapport à une requête particulière
- Évaluer un traitement linguistique « sur site »
  - Permettre une évaluation plus fine des méthodes employées
  - En voir les faiblesses et les avantages en fonction des données traitées
- Pouvoir conjuguer des méthodes spécifiques
  - Adapter le traitement aux données

Quelques efforts dans cette direction

« For most IR algorithms, we do not sufficiently understand the reason for retrieval variability well enough to be able to predict whether the algorithm will succeed or fail on a topic » Buckley & Harman 2003

- Workshop RIA « Where can IR go from here? »
  - 6 semaines de test locaux de plusieurs systèmes
  - Étude de la variabilité requêtes-systèmes-documents
  - Typologie des problèmes rencontrés
    - Termes, relations sémantiques, variations orthographiques, etc.
  - Concentration sur le paramétrage des techniques statistiques
    - Réinjection de pertinence aveugle (blind feedback)
    - Pondération des termes réinjectés
- Pas de lien direct avec des ressources linguistiques

La plateforme RFIEC

- Recherche et Filtrage d'Information et Extraction de Connaissances
  - www.irit.fr/RFIEC
  - C. Chrisment, M. Boughanem, C. Laffaire
- Plateforme de test permettant de :
  - Définir et paramétrer une chaîne complète de RI
  - L'évaluer sur les collections standard
  - Observer le résultat de chacune des étapes
- S'appuie sur :
  - Les collections de tests TREC et CLEF
  - Le moteur de recherche Mercure

Le projet ARIEL

- Adaptation d'une chaîne de Recherche d'Information à l'Expression des besoins sur la base de traitements Linguistiques
  - IRIT / ERSS – J. Mothe – 2005-06
  - Utilisation de la plateforme RFIEC (IRIT)
- Participants :
  - IRIT : N. Aussenac, M. Baziz, M. Boughanem,, C. Laffaire, J. Mothe,
  - ERSS : D. Bourigault, A. Condamines, C. Fabre, N. Hathout, A. Picton, F. Sajous, L. Tanguy, M. Vergez-Couret

Observation des campagnes passées

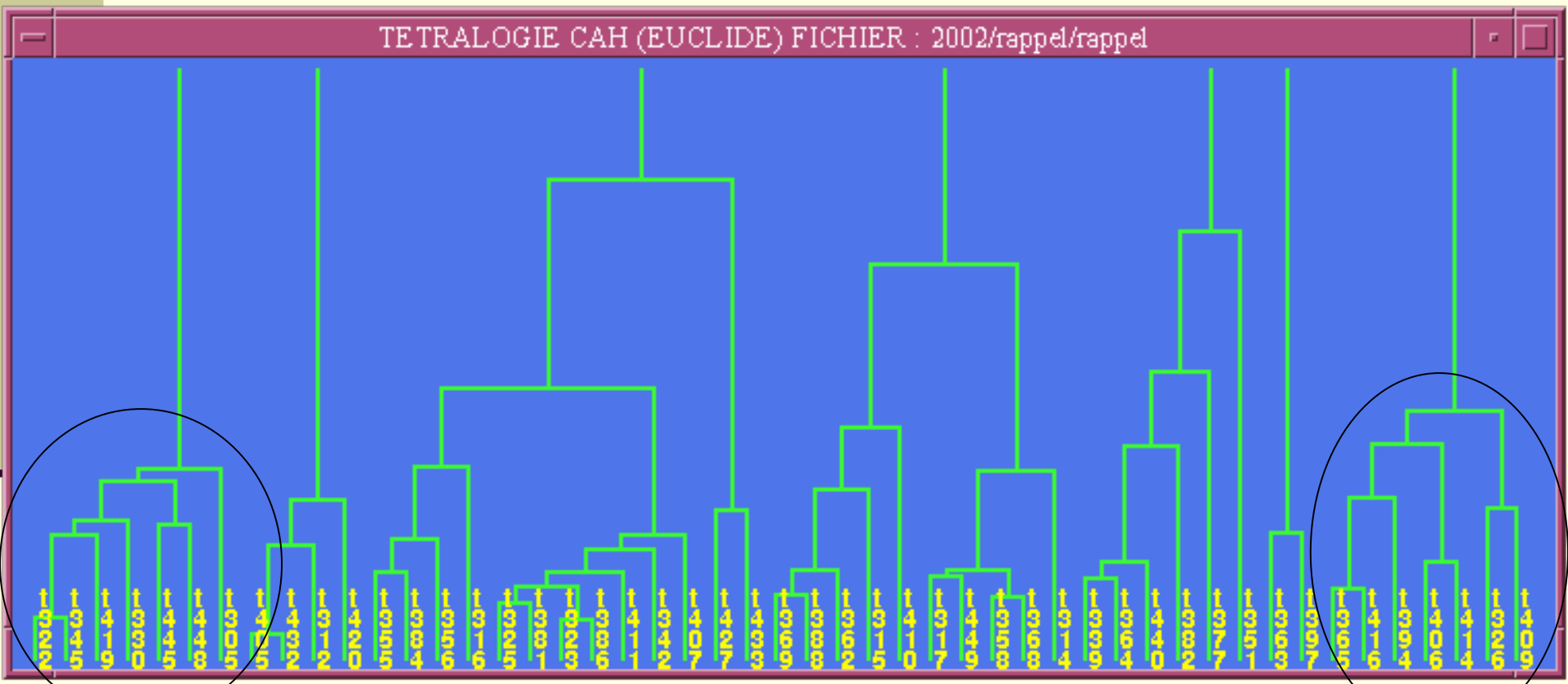
Etude des archives

- Étude des résultats passés des campagnes TREC et CLEF
  - Une masse d'information peu exploitée
- État des lieux des variations de résultats d’une requête à l’autre
- Pour quelques systèmes bien décrits, étude détaillée des différences locales

Typologie naïve des requêtes

- Première approche : requête « faciles » vs « difficiles »
- Étude des runs passés de TREC (5 années, 250 requêtes)
  - Scores de précision, rappel et F-mesure de tous les systèmes et pour toutes les requêtes
- Analyses statistiques
  - Classification hiérarchique ascendante
  - Chrisment et al. (2004)

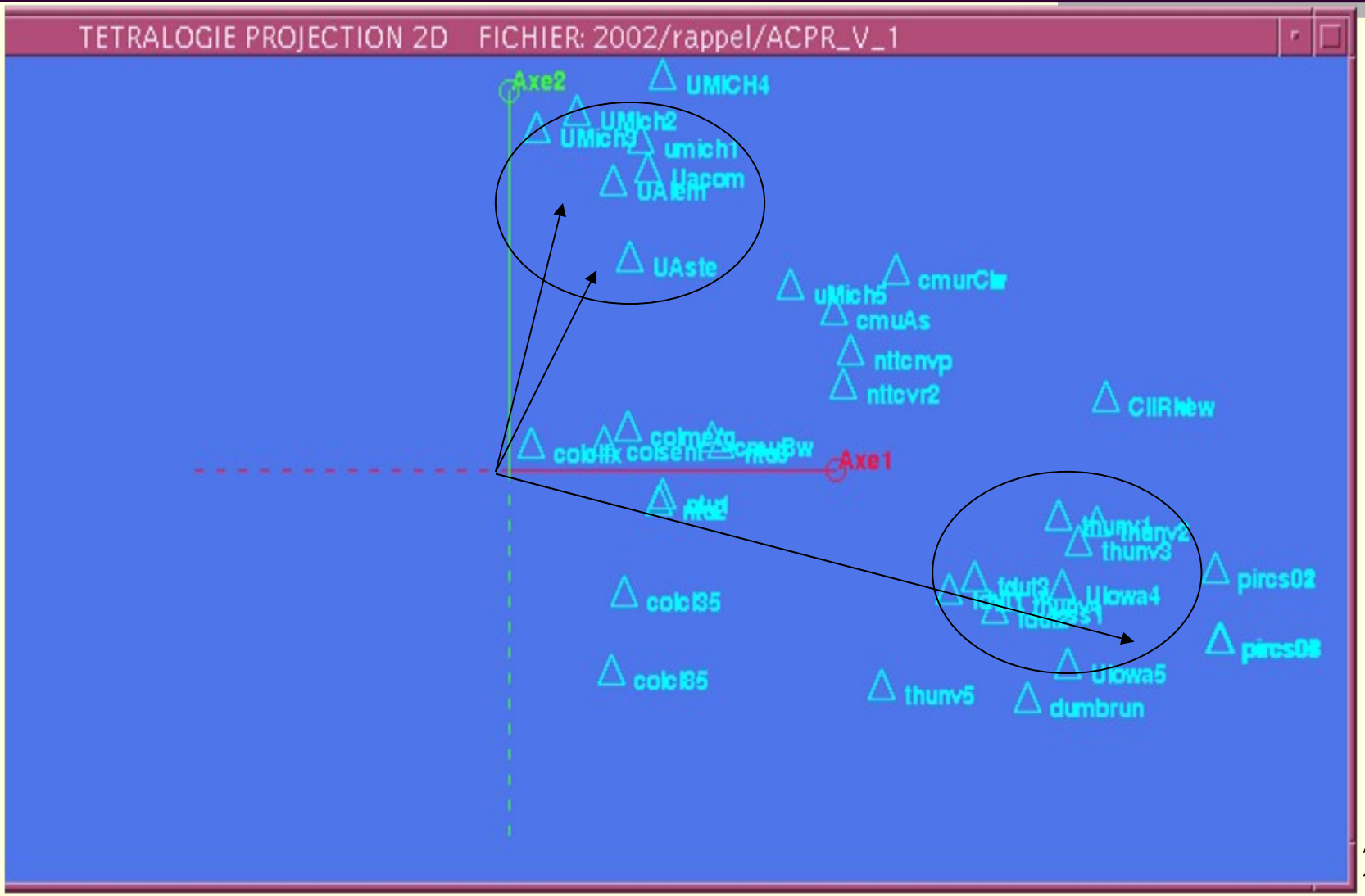
Requêtes faciles vs. difficiles



Typologie des systèmes

- Différences de performances en fonction des requêtes
- Identification de deux groupes de systèmes
- Analyse factorielle
  - Individus = requêtes
  - Variables = scores des runs

Classification des systèmes



Premiers résultats

- Comportements variés entre les requêtes
  - Requêtes difficiles et faciles
- Comportement variés entre les systèmes
  - Efficacité variable d'une requête à l'autre
- Première utilisation : adaptation
  - Pour chaque requête, un type de système est plus approprié
  - Fusion de systèmes : « mixer » les résultats de deux systèmes orthogonaux
- MAIS : pas d' « explication » de la difficulté ni de la variation

Typologie linguistique des requêtes

- Questions :
  - Les caractéristiques linguistiques des requêtes sont-elles liées à leur difficulté ?
  - En fonction de ces caractéristiques, certaines techniques sont-elles plus adaptées ?
- Méthode :
  - Définition de traits linguistiques génériques
  - Corrélation avec la difficulté globale
  - Corrélation avec les variations d'une méthode à l'autre
- Objectif : un système adaptatif

Etudes connexes

- Strzalkowski (1999) :
  - Impact de la longueur des requêtes sur le gain d'une indexation complexe (SN vs mots simples)
  - Les techniques de TAL nécessitent des requêtes longues
- Karlgren (2005) :
  - Etude des caractéristiques stylistiques des textes retrouvés par les systèmes de RI
  - Plus longs, vocabulaire plus riche, plus de mots longs.

Traits linguistiques des requêtes

- Une trentaine de traits définis, répartis en trois niveaux
- Morphologique et lexical :
  - taille des mots (caractères, morphèmes), suffixation, rareté, noms propres, valeurs numériques, répétition (richesse du vocabulaire)
- Syntaxique :
  - nombre de mots, conjonctions, négation, pronoms, complexité syntaxique
- Sémantique :
  - Polysémie des termes de la requête

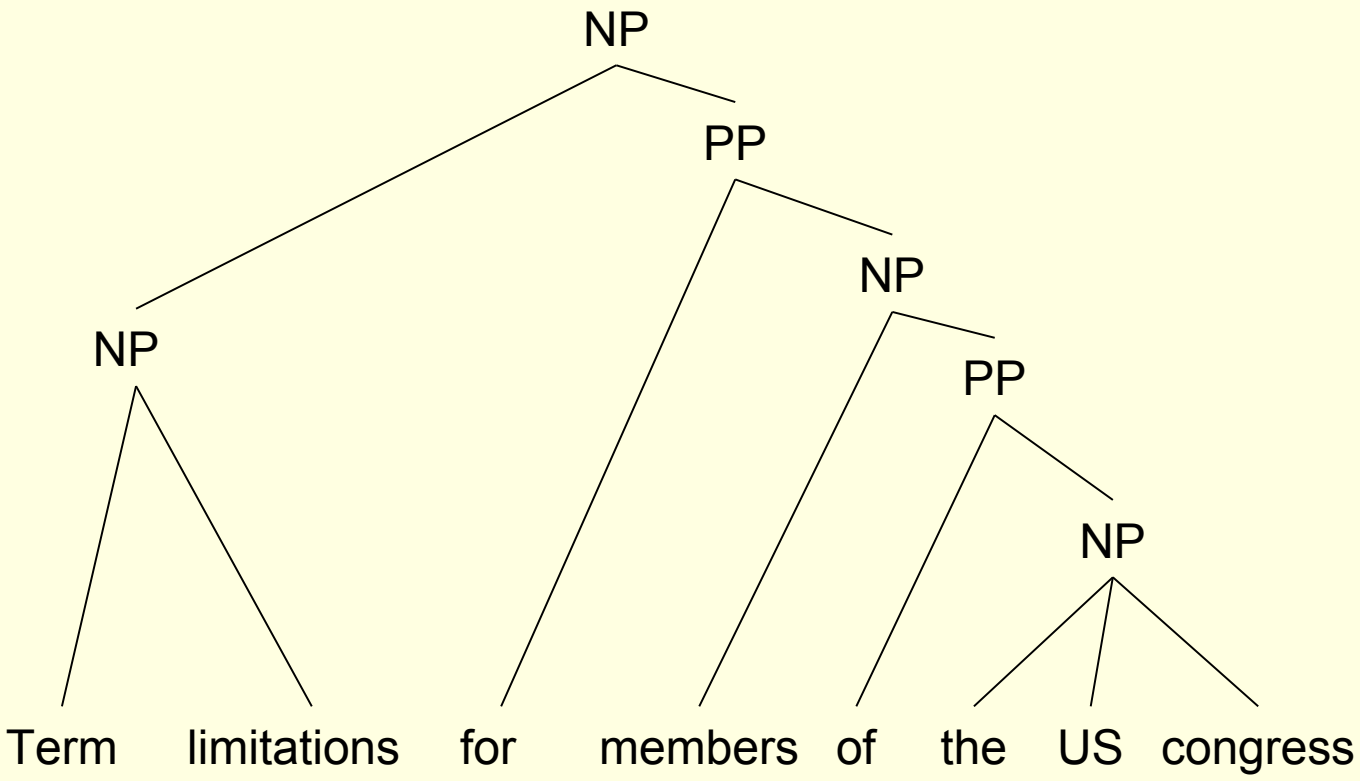
Concrètement…

- Pour chaque requête :
  - Etiquetage morpho-syntaxique (TreeTagger)
  - Analyse syntaxique (Syntex)
  - Projection de ressources génériques
- Profilage :
  - Calculs de caractéristiques (numériques)
  - Construction d'un vecteur pour chaque requête (n=30)
- Grande complexité du problème :
  - Type de tâche et année (adhoc, 5 ans, 50 requêtes/an)
  - Mesures d'évaluation (précision et rappel moyens)
  - Parties de la requête prises en compte (Titre, Titre+Description, Titre+Description+Narration)

Quelques traits particuliers

- Morphologie :
  - Base de suffixes
  - Décomposition en morphèmes par la base CELEX (anglais seulement)
- Sémantique :
  - Anglais : Wordnet
    - Nombre moyen de synsets différents par item lexical de la requête
  - Français : TLFi
    - Nombre moyen d'entrées par item lexical
- Syntaxe : deux approches complémentaires

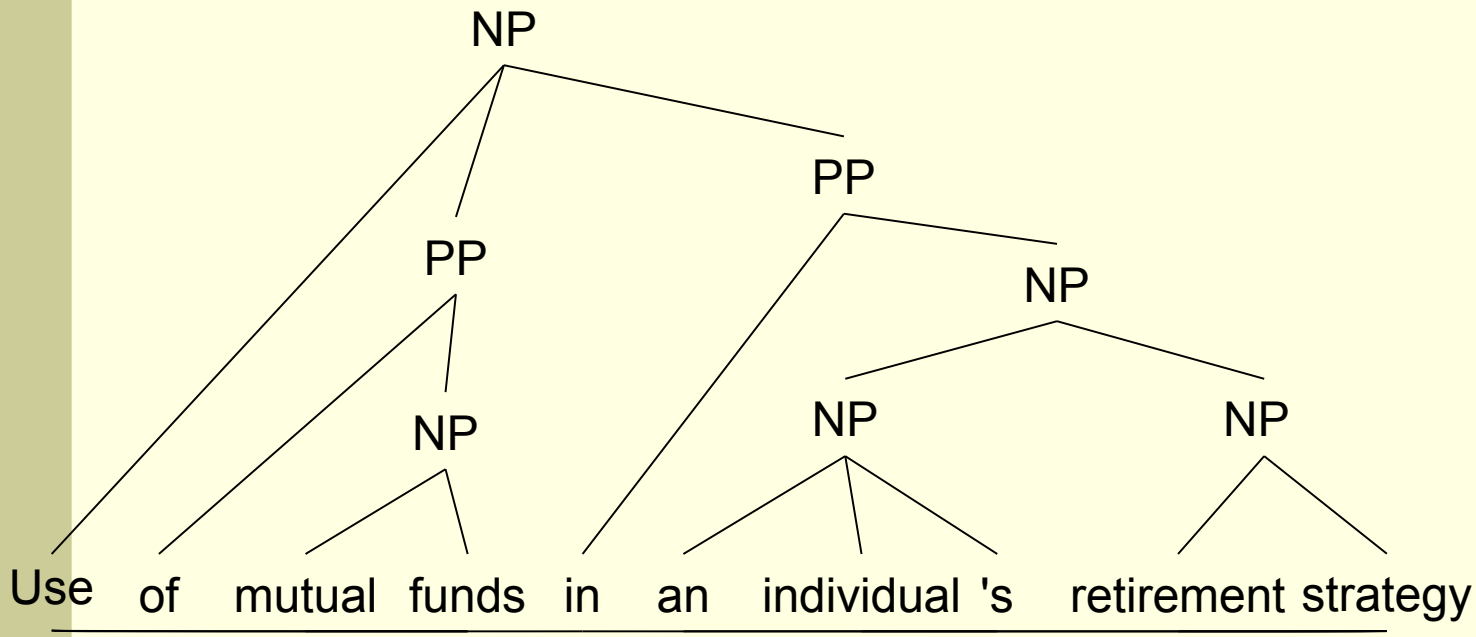
Complexité syntaxique : profondeur



Profondeur
5

Distance
10/7 = 1.43

Complexité syntaxique : distance syntaxique



Profondeur
4

Distance
19/9 = 2.11

Méthode

- Etude des corrélations linéaires (Pearson)
  - p-value < 0,05 et |coefficient| > 0,2
- Corrélation avec le rappel :
  - Négative : polysémie
  - Positive : noms propres
- Corrélation avec la précision :
  - Négative : suffixes, distance syntaxique
  - Positive : noms propres
- Résultats sur les noms propres déjà connus (Mandl & Womser 2004)

Commentaires

- Phénomènes (à peu près) homogènes entre les deux langues
  - (Mothe & Tanguy 2005) : TREC adhoc 1994-1999
  - (Vergez-Couret 2005) : CLEF monoling FR 2000-2005
- Des variations d'une campagne (année) à l'autre
  - Modification des formats, des collections, des consignes
- Permet une prédiction de l'échec
  - Mais pas de façon de lutter contre !
- Encourage vers une exploration des variations des techniques d'une requête à l'autre

Observation locale d'une chaîne de RI

Exploration locale de la chaîne de traitement

- Comparer plusieurs traitements linguistiques
  - Requête par requête
  - Mesurer le gain (ou la perte) en précision
  - Observer les phénomènes linguistiques liés
- Focalisation sur les techniques morphologiques
  - Quatre niveaux de traitements
- Etudier le lien entre les variations d'efficacité et les caractéristiques linguistiques des requêtes

Méthode d'observation

- Plateforme RFIEC : un observatoire indispensable
  - Accès aux données massives
  - Interface de visualisation
- Pour chaque run, pour chaque requête
  - Examen de la requête indexée (quels termes, quelles formes, quelle pondération)
  - Examen de chaque document ramené par le système
    - Pertinent ou non
    - Termes communs avec la requête
- Comparaison de runs
  - Variations dans les documents

Traitements considérés

- Étude sur le français (meilleur contrôle des outils et des ressources)
- Indexation des documents et requêtes
  - Sans traitement
  - Stemmatisation (troncation)
  - Lemmatisation
- Expansion de requêtes
  - Morphologie dérivationnelle
  - (Dictionnaires de synonymes)
  - (Voisins distributionnels)

Traitements morphologiques

- Normalisation des formes :
  - Troncation : « déterminons -> détermi »
    - 5, 6 ou 7 caractères pour le français
  - Lemmatisation : « déterminons -> déterminer »
    - Retour à la forme de citation (flexion)
    - Nécessite un étiquetage morphosyntaxique (TreeTagger)
- Expansion de requêtes suivant les liens de dérivation :
  - déterminer -> détermination
  - Base Verbaction (N. Hathout)
- Etude réalisée sur RFIEC par A. Picton et M. Vergez-Couret (2005-2006)

Normalisation des formes : résultats globaux

- Lemmatisation vs. troncation vs. rien
  - Lemmatisation > Troncation > Rien
  - P5 : (+8%) (+5%)
- Bilan net pour la lemmatisation
  - Variations légères d'une année sur l'autre
  - Variations importantes entre les requêtes

Lemmatisation vs Troncation : explications locales

- Flexion de mots courts
  - loi/lois, élire/élit
  - +Lemm, -Tronc
- Formes dérivées
  - ménopause / ménopausique
  - +Tronc –Lemm
- Formes apparentées non pertinentes
  - information / informatique, fonctionnement / fonctionnaire
  - -Tronc, +Lemm
- Problèmes de lemmatisation
  - Adjectif vs participe passé :
    - accouché (Adj) / accoucher (V)
  - Erreurs de lemmatisation dûes à l'étiquetage :
    - Traité (de paix) -> traiter

Traits linguistiques liés à la variation entre lemmatisation et troncation

- Corrélation positive avec un gain en faveur de la lemmatisation :
  - Longueur de la requête (phénomène connu)
  - Nombre de mots suffixés
  - Noms propres
- Danger de la dérivation non contrôlée :
  - Information / informatique
- Danger des noms propres :
  - Hollande / hollandais

Expansion morphologique de requêtes

- Base Verbaction (Hathout et al.)
  - 9000 couples nom/verbe validés manuellement
  - Le nom et le verbe sont morphologiquement liés
  - Le nom dénote l'action exprimée par le verbe
  - Ressource obtenue à partir du TLF et complétée par acquisition sur le Web
- Intervention lors de l'indexation de la requête après lemmatisation
  - Double sens possible : N -> V et V -> N
- Environ 51% des relations de Verbaction ne sont pas retrouvables par troncation

Bilan global

- (Lemmatisation + Verbaction) vs. Lemmatisation seule
  - Amélioration très faible
- (Lemmatisation + Verbaction) vs Troncation
  - Amélioration un peu plus sensible
- Pas de résultat net (variations entre les années et les mesures de précision)

Phénomènes linguistiques locaux

- Bruit induit par Verbaction
  - Problèmes de catégorisation : *égaler/également* (N vs. Adv)
  - Polysémie des dérivés : *faire/faction, faire/facture, faire/façon*
- Gain de Verbaction :
  - Bons couples dans les deux sens (N-V et V-N)
  - *Investir/investissement, adhésion/adhérer, union/unir*, etc.
- Insuffisance de Verbaction (vs. troncation) :
  - Liens morphologiques autres que N/V : *ménopausique/ménopause* (Adj/N)
- Compensation des erreurs d'étiquetage :
  - (effet de) serre -(lemm.) -> serrer -(exp.)-> serre

Traits linguistiques corrélés

- Traits positivement corrélés avec un gain de Verbaction sur la lemmatisation seule :
  - Prépositions (+)
  - Noms (+)
- Traits positivement corrélés avec un gain de Verbaction sur la troncation :
  - Noms (+)
  - Noms Propres (-)
- Les meilleures expansions se font du Nom vers le Verbe

Autres phénomènes à prendre en compte

- Pondération des termes ajoutés lors de l'expansion
  - Poids égal ou inférieur à celui des termes initiaux
  - Cumul des poids lors des répétitions (notamment des flexions d'un même verbe)

Expérimentations et conclusion

Utilisation des résultats

- Proposition d'un système adaptatif à la campagne CLEF 2006 (Y. Loiseau)
  - Tâche monolingue français
  - Trois propositions :
- 1/ Application de la méthode la plus adaptée en fonction des caractéristiques des requêtes
  - Arbre de décision sur les caractéristiques linguistiques
  - Appris sur les 6 campagnes précédentes
- 2/ Fusion des 4 méthodes envisagées
  - Union des 4 listes obtenues pour une requête
- 3/ Baseline : troncation simple

Et le vainqueur est…

- L'université de Neuchâtel…
- Parmi nos trois propositions (27 participants)
  - Fusion : 8è
  - Système adaptatif : 11è
  - Troncation simple : 14è

Bilan : points positifs

- Démonstration du gain de méthodes linguistiques
  - Méthodes simples et robustes
  - Bien contrôlées et articulées
  - Avantage aux ressources limitées mais validées
    - Encouragement pour leur développement
    - Extension de la couverture

Bilan : points négatifs

- Pas de liaison fiable entre les caractéristiques linguistiques et les méthodes
  - Trop grande variation d'une année à l'autre
  - Variation en fonction de la mesure utilisée
- Méthode brutale encore gagnante
  - Fusion "aveugle" plus rentable qu' "éclairée"

Perspectives

- Affinement des traits descriptifs
- Étude des documents et pas seulement des requêtes
- Contrôle local des techniques