Le Web comme source de connaissances pour améliorer la fiabilité des réponses

Groupe LIR - LIMSI (Orsay)

Brigitte Grau, Isabelle Robba, Anne Vilnat

Problématique

- **Réponse à des questions factuelles**
- **La réponse exacte et un score de fiabilité**
  - What is the name of the volcano that destroyed the ancient city of Pompeii?
    - (Mount) Vesuvius
  - How many chromosomes does a human zygote have?
    - 46 ou 23 pairs
  - What lays blue eggs?
    - Araucana (hen)
  - What was the first spaceship on the moon?
    - Eagle
  - How did Mahatma Gandhi die?
    - shot dead

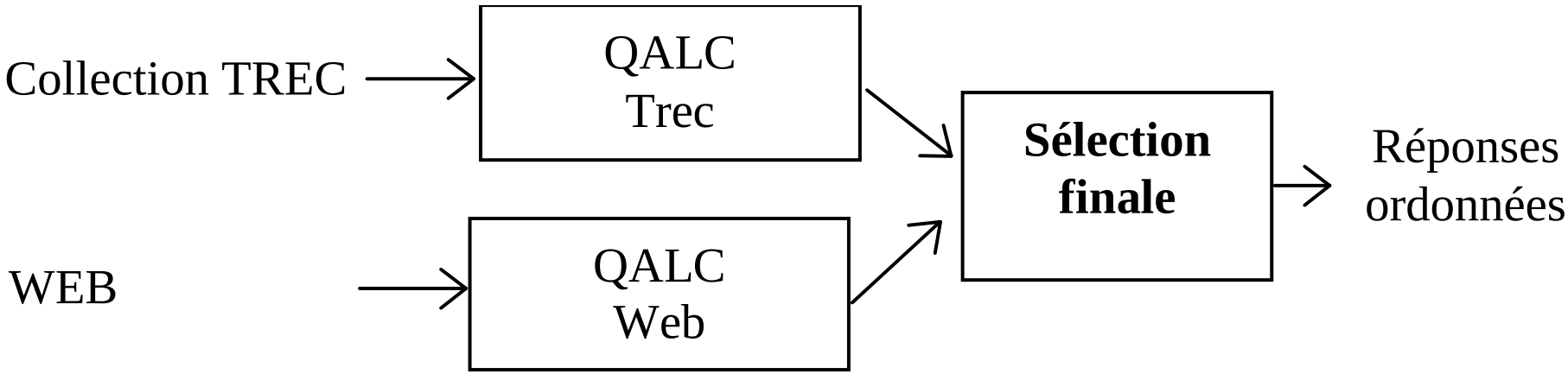
Caractéristiques

- **Questions plus ou moins difficiles selon :**
  - La diversité dans la formulation des réponses
    - Dans les termes utilisés
    - Dans le style
  - La redondance des réponses
    - Confirmation par différentes sources
- **Collection fermée d'articles de journaux**
  - Sources fiables, pb de manque de diversité et redondance
- **Web**
  - Diversité et redondance, pb de fiabilité

Différentes approches (Trec 11)

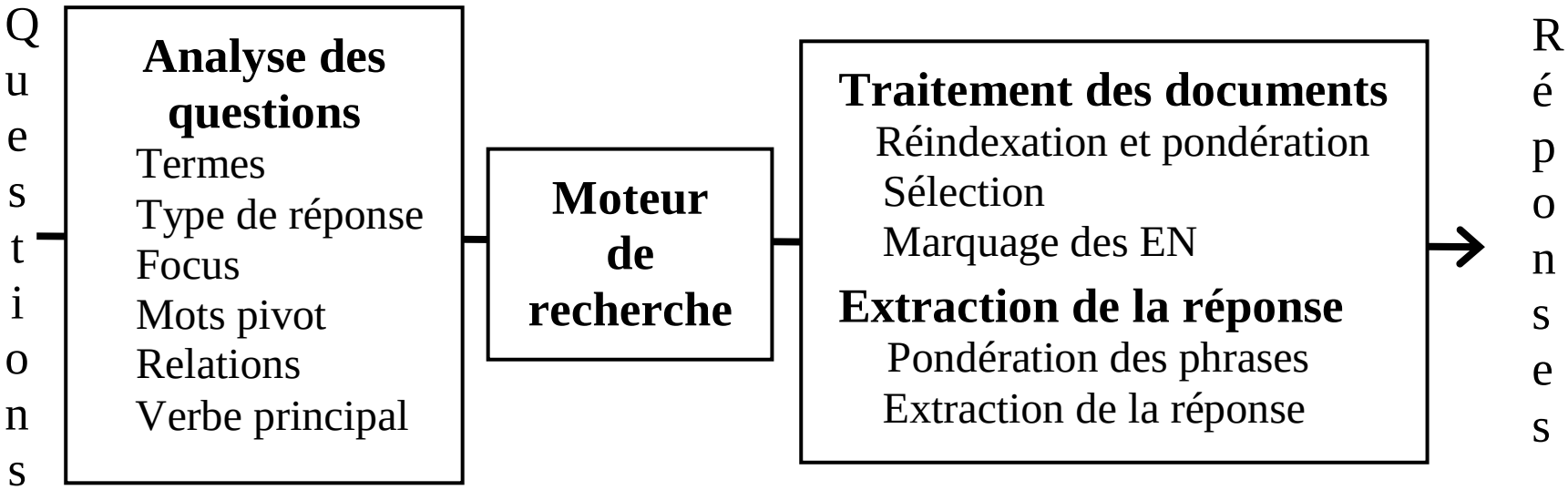
- **Confirmer une réponse par le Web**
  - Magnini et al. : + 28%
  - Requête booleenne avec mots de la question + réponse
  - Validité de la réponse fonction du nombre de documents trouvés
- **Augmenter le facteur de redondance**
  - Clarke et al. : + 25 à 30%
  - 40 premiers documents trouvés par 2 moteurs
  - 20 premiers passages collection
- **Utiliser uniquement le Web**
  - Brill et al.

Deux avis valent mieux qu'un …



- **Recherches en parallèle**
  - 5 meilleures réponses
- **Sélection finale**
  - La même réponse figurant dans chacune des listes

Le système QALC



- **Variantes entre collection fermée et Web**
  - Formulation des requêtes
  - Nombre de documents retenus

Requête Web

- **Utilisation de Google**
- **Formulation précise de la réponse**
  - La plus courante
  - Etude manuelle sur 50 questions
- **Reformulation de la question à l'affirmative**
  - Fondée sur les caractéristiques des questions
    - Type de réponse attendu
      - Who is the governor of Alaska : PERSONNE
      - Where is the Devil's Tower : LIEU
    - Type de question : forme syntaxique abrégée
    - Principales caractéristiques : Focus, verbe, relations syntaxiques, préposition

Analyse de la Question : méthode

- **Analyse syntaxique (analyseur IFSP de Xerox):**
  - Repérage des types d'interrogateurs
    - who : personne ou institution
    - how much : quantité numérique
  - Délimitation des groupes (verbes et groupes nominaux)
- **Critères sémantiques**
  - Définition de classes / types d'entités nommées
- **Exemples de règles :**

HOW MUCH Auxiliaire GN VerbeFinancier :

➔ MONTANT-FINANCIER

WHAT VerbeEtre GNNameOf GNFonction :

➔ PERSON

Analyse des questions

Règles pour les types sémantiques

- **Selon le nom tête du Groupe Nominal sur lequel porte l'interrogatif :**
  - What flower did Vincent Van Gogh paint ?
    - FLOWER
  - Name an art gallery in New Work ?
    - GALLERY
- **Doit appartenir à une base de connaissances**
  - WordNet
- **Pas de type attendu de la réponse**
  - What killed Bob Marley ? => ????

Analyse des questions

Règles pour déterminer le focus

- **Focus :**
  - Souvent le sujet syntaxique de la question
    - Le mot tête du GN
    - Plus les modifieurs
- **Définition de règles**
  - What kind of GN1 be GNSUJET .. ?
    - ⇨ *What kind of animal was Winnie the Pooh ?*
  - What Be the name of GN … ?
    - ⇨ *What is the name of the second space shuttle ?*
  - How ADJ be GN … ?
    - ⇨ How old is the sun ?
  - How many GN verbe GNOBJ .. ?
    - ⇨ How many states have a lottery ?

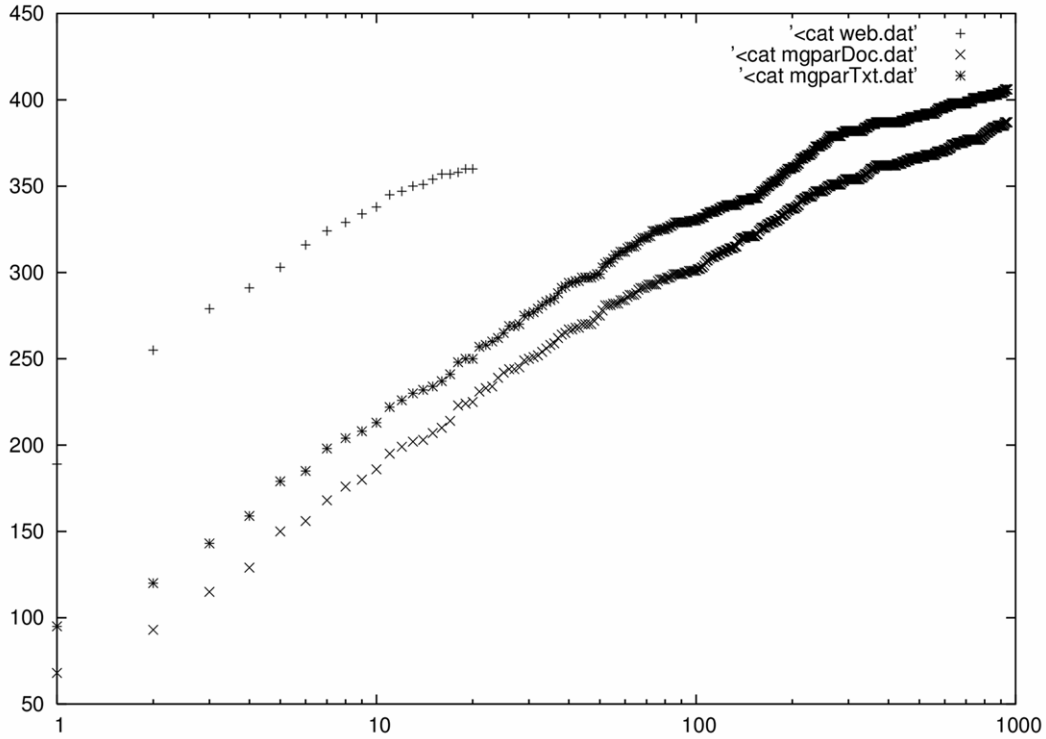
Formulation de la requête Web

- **Schémas de réécriture :**
  - DATE : <focus> <verbe principal> born on
    - When was Lyndon B. Johnson born?
    - Lyndon B. Johnson was born on
- **Terme introduisant un modifieur**
  - ex. conjonction pour une circonstancielle
- **Terme introduisant la réponse / type de réponse**
  - ex. in pour un LIEU, on pour une Date
- **Schéma par défaut**
  - Tous les mots sans le pronom et l'auxiliaire interrogatifs
- **Requêtes**
  - Application du ou des schémas avec et sans guillemets

Recherche des documents

- **Web**
  - 20 premiers documents
- **Collection AQUAINT**
  - Caractéristiques
    - 1 million d'articles de journaux (3 gigas)
    - Collection découpée en paragraphes
      - Indexation par MG, utilisation du stemming
  - Requête « lâche »
    - Mots pleins de la question, lemmatisés et non lemmatisés
    - Focus
  - Liste ordonnée de documents (cosinus)
    - 1500 paragraphes retenus

Nombre de documents retenus



- **20 documents trouvés par Google**
- **1500 paragraphes trouvés par MG**

Extraction de la réponse

- **Pondération des phrases**
  - Mots et termes présents (tels quels ou variations)
  - Proximité des termes
  - Présence de l'entité nommée attendue
- **Extraction de la réponse**
  - L'entité nommée (le type attendu ou un type plus général)

ou

  - Application de patrons d'extractions / type de question
  - Augmentation du poids de la phrase
  - Possibilité de réponse NIL : pas de réponse
  - Conserve les 5 premières réponses + poids

Résultats

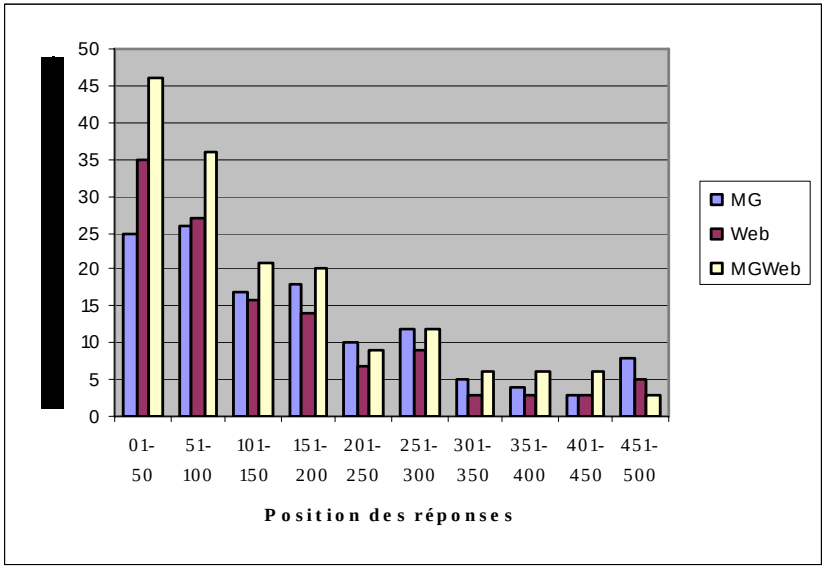
- **Participation à Trec 11 (2002)**
- **500 questions**
  - 312 catégorisées par QALC comme EN attendue (8 erreurs)
- **500 réponses classées / degré de confiance**
- **Mesure** $\frac{1}{Q}\sum_{i=1}^{Q}\frac{\text{Nb réponses correctes dans les i premiers rangs}}{i}$
- **Résultats de QALC :**
  - 133 réponses exactes, 11 inexactes, 20 non justifiées
  - Score : 0.497
  - 9ème/34

Analyse des résultats

- **A partir des patrons de réponse fournis par le NIST**
- **Plus laxiste**
  - Inexactes et non justifiées sont incluses (~+30)
- **Permet la comparaison**

	Bonnes réponses	Score
AQUAINT	**128**	**0.402**
Web	**122 (-0,04%)**	**0.436 (+0,08%)**
AQUAINT +Web	**165 (+29%)**	**0.587 (+46%)**

Amélioration de la fiabilité des réponses



- 106 réponses communes (rangs 1 à 5)
- 42 dans collection AQUAINT
- 17 sur le Web
- Meilleur classement Web dans 100 premiers
- Excellent classement avec combinaison

Conclusion

- **Fiabilité : selon les résultats identiques issus de différentes sources**
- **Recherche Web :**
  - Requête précise possible
  - Uniquement liée à la formulation de la question (peu de variation)
  - Peu de documents à analyser
- **Recherche collection fermée**
  - Requête large, car seules les variations dérivationnelle et morphologique gérées à ce niveau
    - Ramener aussi les documents avec synonymes
  - Sélection des passages et extraction des réponses : gérer la variation

Perspectives

- **Stratégies d'extraction propres au Web**
  - En liaison avec la formulation de la requête
- **Passage au français**
  - Evaluation EQueR (projet Evalda - Technolangue)
- **Vérifier que le Web français a le même comportement**
  - Redondance ?
  - Diversité ?