Evaluation et systèmes complexes de recherche d'information

Brigitte Grau
LIMSI – CNRS (Orsay)

Que signifie évaluer ?

- «Action d'évaluer, d'apprécier la valeur (d'une chose); technique, méthode d'estimation.» (TLFi)
- Démarche scientifique + technologique
  - Important pour estimer le succès d'une recherche
- Appréciation qualitative ou quantitative de la performance d'un système, avec comme point de référence la performance humaine.

Pourquoi évaluer ?

équipes de R&D

•valider hypothèses de recherche
•confrontation des résultats entre équipes de recherche
•définition de tâches communes, construction de référentiels, clarification de terminologie

acteurs industriels

• identifier technologies prometteuses, décider si technologie suffisamment mature et robuste pour application commerciale

agences de financement

•mesurer avancées technologiques

•clarifier l'offre technologique

utilisateurs

Le paradigme d'évaluation

Évaluation

=

expériences

+

comparaisons

Le paradigme d'évaluation

- Définir un champ expérimental commun
  - Données
  - Représentations
  - Mesures
- pour des problèmes universels
- et en dégager des résultats reproductibles

Critères : intrinsèque/extrinsèque

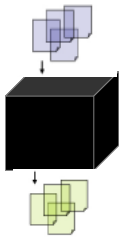
- Intrinsèque : système évalué par rapport à une référence (gold standard)
- Extrinsèque ou évaluation en contexte : évaluation dans un système complet répondant à une fonction précise pour l'utilisateur
- Exemple pour un analyseur syntaxique :
  - intrinsèque : justesse des résultats d'analyse comparés à ce qui était attendu
  - extrinsèque : impact des résultats dans un système de questions-réponses

Evaluation automatique/manuelle

- Automatique : comparaison à la référence
  - production de la référence (coûteuse; guide d'annotation nécessaire) puis évaluation aisée
  - impossible dans certains domaines (ex : traduction automatique)
- Manuelle
  - Coûteux
  - Non réutilisable pour la comparaison
- Problème de l'accord inter-annotateur
  - très bon sur étiquetage grammatical (POS) par exemple (>90%)
  - désambiguïsation : de l'ordre de 60% ($\kappa \approx 0.3$) (Yong, 1999)...
  - adjudication éventuelle

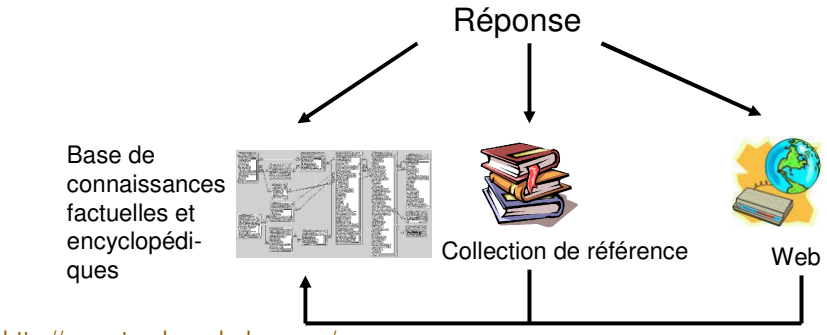
Méthode d'évaluation

- **Boîte noire** : sorties pour une entrée donnée + éventuellement évaluation de performance, i.e. vitesse, fiabilité, ressources etc.
  - Cas des campagnes d'évaluation
- Application à la recherche d'information précise
  - Système de question-réponse (QR)



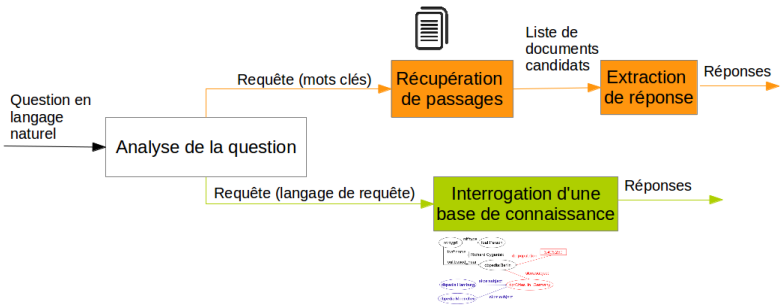
Système de question-réponse

- A partir d'une requête en langue naturelle
  - car plus précis qu'une requête sous forme de mots-clef
    - Quand Sangatte a-t-il été créé ? / Sangatte, créer



http://www.trueknowledge.com/
http://www.wolframalpha.com/

Système de question-réponse

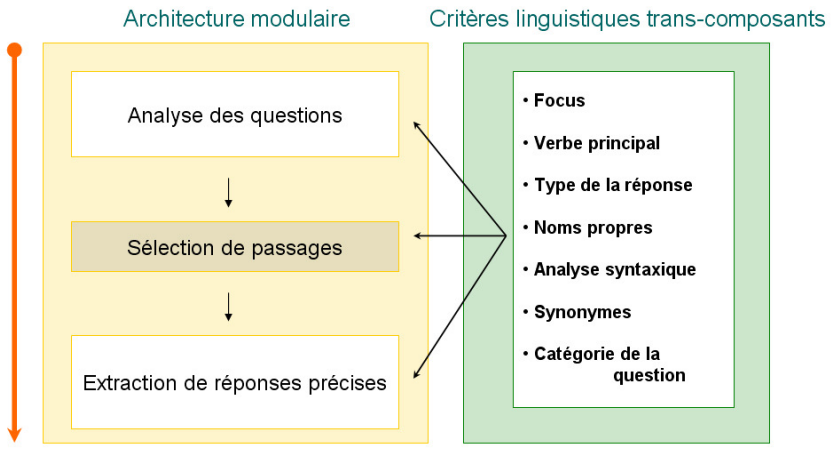


- Texte
  - × Non structuré
  - × Variations et ambigüités
  - ✓ Nombreuses informations
- Base de connaissances
  - ✓ Structurée
  - ✓ Normalisée, non ambigüe
  - × Informations manquantes

Evaluation de QR

- Sur du texte
  - TREC, QA@CLEF, EQuER, Quaero, NTCIR
  - QCM & Machine reading : QA4MRE@CLEF
- Sur des bases de connaissances
  - QALD, BioASK @ CLEF
    - DBpedia, KB médicale, etc
- Mesures
  - 1 réponse fournie par question : proportion de réponses correctes (accuracy)
  - N réponses par questions : MRR (Mean Reciprocal Rank)
  - Variantes de l'accuracy pour tenir compte de l'absence de réponse, du degré de confiance du système en la réponse

QR dans du texte



Campagnes d'évaluation

- Que permettent les campagnes ?
  - Connaître la performance globale des systèmes
    - Théoriquement : savoir quel est le meilleur système (mais …)
  - Disposer de corpus
    - Pour évaluer l'amélioration des systèmes
    - Connaître l'intérêt d'une ressource
      - Evaluation par ablation
    - Pour entraîner les systèmes (mais …)
- Corpus
  - Corpus construits par les organisateurs
    - Questions (200 to 500): factuelle, définition, complexe, booléenne
    - Plus ou moins indépendamment des documents
  - Corpus construits à partir des runs des systèmes participants :
    - Réponse précise : court extrait de texte, annotée juste ou faux
    - Passages supports (souvent 250 caractères) contenant la réponse extraite
      - Pour aider l'évaluation humaine des résultats

Campagnes d'évaluation : Corpus

<q q_id="239" q_time="0.176134">
<q_str> **who directed the film Rough South of Harry Crews ?**</q_str>
<a a_id="239_1" type="text" complexity="**S**" boolean="**Nothing**" rank="**1**" doc_id="**00172736.0**"
status="**Supported**" evaluation="**false**" >
<a_short_str> **Dennis Miller Show** </a_short_str>
<support> **won a 1992 regional Emmy award for The Rough South of Harry Crews , commissioned by UNC-TV . Hawkins also directed a film about Southern author Larry Brown , The Rough South of Larry Brown . the Dennis Miller Show** </support>

<a a_id="239_2" type="text" complexity="**S**" boolean="**Nothing**" rank="**2**" doc_id="**00172736.0**"
status="**Supported**" evaluation="**true**" >
<a_short_str> **Gary Hawkins** </a_short_str>
<support> **Gary Hawkins , who currently teaches at Duke University , won a 1992 regional Emmy award for The Rough South of Harry Crews , commissioned by UNC-TV . Hawkins also directed a film about Southern author Larry Brown , The Rough South of Larry Brown .**</support>

</q>

- Corpus bruts sans annotation
  - Question: type de réponse attendue, structure syntaxique, focus, etc …
  - Passage: entité nommée, structure syntaxique, termes de la question, etc.

Campagnes d'évaluation : limites

- Ce que ne permettent pas les évaluations boites noires
  - Savoir pourquoi une méthode est meilleure
    - Quels sont les phénomènes traités ?
    - ➢ Pas de classification des questions *a priori*, car dépendante du texte
    - Dépendances aux outils / Impact des erreurs dues aux différents modules

Corpus des campagnes d'évaluation

- De quoi les corpus sont-ils représentatifs ?
  - De la tâche de recherche de réponses
  - Des différentes formes de questions
    - Différents types de réponses
    - Différentes formulations
- De quoi les corpus ne sont-ils pas représentatifs ?
  - Difficulté de trouver une réponse
    - Beaucoup de questions sans réponses → absentes des corpus
    - Quand la réponse existe : souvent, une reformulation directe de la question
    - → Absence des propriétés linguistiques permettant de trouver des équivalences entre question et réponse :
      - Synonymie
      - Paraphrase: syntaxique and sémantique

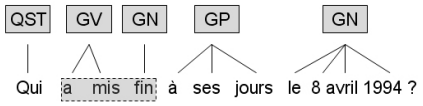
Pour améliorer les systèmes

- Analyses d'erreurs
  - Travail individuel
  - Pas de méthodologie commune, pas d'outils partagés
- Evaluation modulaire
  - Intrinsèque :
    - Sans problème si module = tâche en elle-même (EN, POS)
    - Sinon : définir un cadre d'évaluation
  - Extrinsèque : lourd
    - Pouvoir tracer les erreurs
    - Indirecte en remplaçant un module par un autre

Analyse d'erreurs

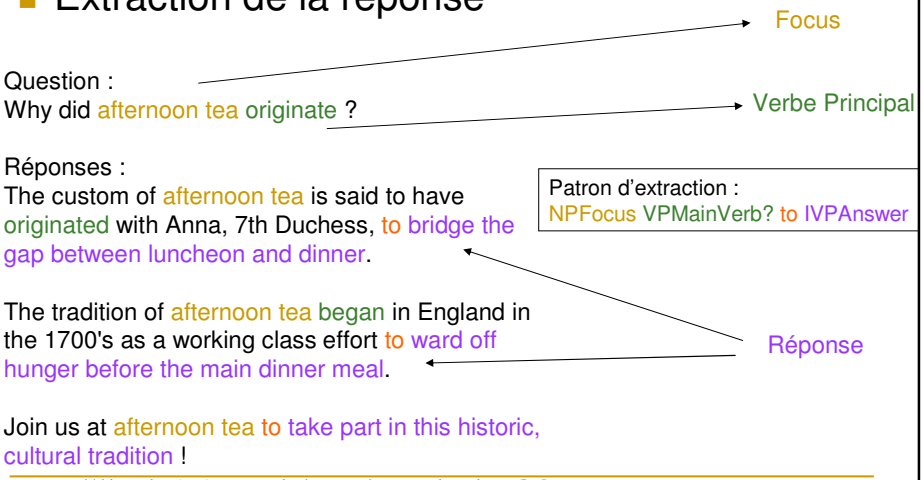
- Analyse de la question
  - Etiquetage morphosyntaxique (POS)
  - Analyse syntaxique
  - Extraction de critères :
    - Où l'A 340 a-t-il établi le record du plus long vol sans escale ?
    - Focus: A

Forme	Etiq	Lemme
Dans	IN	dans
quelle	DT	quel
ville	NN	ville
chinoise	VVZ	chinoiser
le	DT	le
secrétaire	NN	secrétaire
d'	IN	de
Etat	NN	état
américain	JJ	américain
Warren	NP	Warren
Christopher	NP	Christopher
a	VHZ	avoir
-t-il	PP	il
été	VBN	être
mandaté	NN	mandaté
par	IN	par
Bill	NP	Bill
Clinton	NP	Clinton
?	SENT	?



Analyse d'erreurs

- Extraction de la réponse

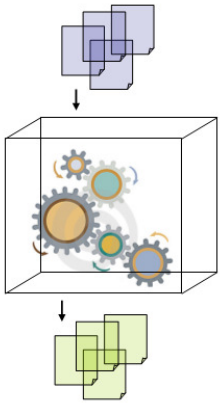


Evaluation modulaire

- Évaluation modulaire : division en sous-tâches
  - Entités Nommées : MUC, ACE
  - Etiquetage morpho-syntaxique (GRACE)
  - Textual Entailment (RTE)
  - …
- Avantage
  - Évaluation de phénomènes linguistiques en dehors de leur cadre d'application (hors contexte).
- Limites
  - Réintroduction de ces outils en contexte augmente l'impact des erreurs récurrentes
  - Difficultés sur des corpus spécifiques
  - Impact sur les performances globales difficiles à évaluer

Evaluation transparente

- **Boîte transparente** : évaluation de la structure interne
  - Composants du système, ressources linguistiques utilisées, phénomènes linguistiques traités...
  - Suppose un accès aux résultats produits au sein du processus global



→ Thèse Sarra El Ayari, 2009, LIMSI

Evaluation transparente

- Résultats intermédiaires → corpus
  - Sélection des documents avec la réponse correcte dans la requête
  - Le système de QA pour produire des annotations automatiques
  - Possibilité de compléter manuellement
- Evaluation transversale des propriétés linguistiques :
  - Tracées dans tout le système
    - Après application des différents modules
  - Propriétés étudiées surlignées
  - Typage des erreurs par ajout d'annotations
  - Correction manuelle des erreurs et évaluation de l'impact sur les performances
    - Application du système sur des données corrigées

Evaluation transparente

173) En combien de provinces est divisé l' Afghanistan ?

- **Réponses correctes : 30 trente**
- **Verbe principal : diviser**
- **Focus : Afghanistan**

- **268 phrases réponses contiennent le focus**
- **8 phrases réponses contiennent le focus et la réponse**

1	Par ailleurs , des forces de Massoud et de Dostam continuaient de se affronter dans plusieurs provinces du Nord de le Afghanistan **, les deux camps se attribuant des victoires .**	0
2	Les taliban , qui contrôlent neuf provinces du sud de le Afghanistan **, et les chiites du Wahdat (faction pro-iranienne) ont été chassés de Kaboul durant le week-end par une offensive éclair de Massoud .**	0
3	Ils avaient volé de victoire en victoire ces six derniers mois , prenant le contrôle de neuf provinces de le Afghanistan **(sur trente) avant de parvenir à la mi-février aux portes de Kaboul .**	1
4	Troisième journée de combats à Kaboul Extension des affrontements au provinces du nord de le Afghanistan **synthèse . Kaboul/ Islamabad , 3 jan (ats/ afp/ reuter)**	0
5	Les Talibans , qui ont surgi sur la scène politique afghane à le automne dernier , avec la prise de Kandahar , le ancienne capitale royale , ne ont cessé depuis lors de gagner du terrain dans le sud de le Afghanistan **où ils contrôlent huit provinces .**	0
6	Si la chute du Logar se confirmait , le mouvement , surgi à le automne dernier à la frontière pakistanaise , au sud , contrôlerait à présent neuf provinces de le Afghanistan **sur 30 .**	1
7	La radio gouvernementale a affirmé , mercredi 4 janvier , que des avions russes et de autres pays de la Communauté des Etats indépendants (CEI) avaient récemment bombardé deux des provinces du nord de le Afghanistan **limitrophes du Tadjikistan , tuant une dizaine de civils et détruisant des bâtiments .**	0
8	Entrés dans le " grand jeu " afg-han à le automne 1994 , les talibans de jeunes guerriers aux convictions intégristes qui se étaient d'abord réunis autour de écoles coraniques , notamment dans les camps de réfugiés du Pakistan avaient rapidement accumulé les succès , jusque à dominer , à la fin de le hiver , une douzaine de provinces du sud de le Afghanistan **, la seule partie du pays qui , à ce jour , demeurait sans structure politique après la chute , en avril 1992 , du communisme .**	0
9	le avancée des talibans , les " étudiants religieux " qui sont parvenus aux portes de Kaboul le mois dernier , après avoir conquis neuf provinces du sud de le Afghanistan **, a mis le Wahdat dans une situation militaire très précaire .**	0

QR à CLEF

http://nlp.uned.es/clef-qa/

- QA4MRE@CLEF : Question Answering for Machine Reading Exercice
- Un texte, un QCM de ~ 20 questions / 4 ou 5 réponses, dont une seule est correcte
- Textes :
  - articles scientifiques de domaines différents
  - textes narratifs : examens d'entrées
- Intérêt :
  - Maitriser les difficultés à résoudre : le texte support est fixe

QA4MRE : examen d'entrée

The white-haired old man was sitting in his favorite chair, holding a thick book and rubbing his tired eyes. When his nineteen-year-old granddaughter, Valerie, came into the room, he looked up and smiled. His eyes instantly brightened with happiness to see her. 'Hi, Grandpa. What are you reading?' she asked, pulling up a chair beside him. 'Oh, it's a book on the architecture of Spain. But I'm not really reading. Mostly I am just falling asleep over the pictures,' he said, laughing. 'Are you finished packing your bags yet?' he asked. The following morning Valerie and two of her friends were flying to Europe for a two-week holiday. 'Almost. I need to travel light, you see, so I can buy lots of new dresses and shoes in Paris and Barcelona.' They both laughed because Valerie was not actually interested in fashion at all. She loved foreign languages, music, art, good food, and many other things - but not shopping for clothes. […]

<question q_id="1">
<q_str>Why did Valerie and Grandpa laugh?</q_str>
<answer a_id="1">Valerie had not finished her preparation.</answer>
<answer a_id="2">Valerie had too many things in her suitcase.</answer>
<answer a_id="3">They both knew that what Valerie said was not true.</answer>
<answer a_id="4">They both understood that Valerie had very little money.</answer>
</question>

QR dans des bases de connaissances

QALD @ CLEF

- QALD : tâche 1 multilingue
  - Questions avec agrégat / fonction
    - Combien de fois s'est mariée Jane Fonda?
    - SELECT COUNT(DISTINCT ?uri) WHERE { res:Jane_Fonda dbo:spouse ?uri . }
  - Questions listes
    - Quels livres de Kerouac ont été publiés par Viking Press?
    - SELECT DISTINCT ?uri
    - WHERE {
    - ?uri rdf:type dbo:Book .
    - ?uri dbo:publisher res:Viking_Press .
    - ?uri dbo:author res:Jack_Kerouac .
    - }
  - Questions complexes
    - Quel état des Etats-Unis a la plus haute densité de population?
    - SELECT DISTINCT ?uri
    - WHERE {
    - ?uri rdf:type yago:StatesOfTheUnitedStates .
    - ?uri dbp:densityrank ?rank .
    - }
    - ORDER BY ASC(?rank)
    - OFFSET 0 LIMIT 1

QALD @ CLEF

- Problèmes
  - Déterminer les termes désignant les entités, relations, catégories
  - Déterminer ces notions dans la KB : variations linguistiques
  - Construire la requête SPARQL
- Tâche 1 : 50 questions
- Tâche 2 (médical) : 25 questions
- Est-ce suffisant pour représenter la tâche ?

Conclusion

- Le paradigme d'évaluation : trois questions ouvertes
  - Quelle est la taille nécessaire des données de référence ?
  - Quel est le niveau minimum de qualité d'annotation de la référence ?
  - Comment faire des annotations cohérentes sur de grand volumes de données à faible coût ?
    - Croudsourcing ?

Conclusion

- Campagnes d'évaluation
  - Avantages :
    - Tâches évolutives (différentes chaque année)
    - Constitution de corpus d'étude
    - Métriques d'évaluation de référence
  - Difficultés
    - Tâches évolutives (différentes chaque année)
    - Définition de la tâche à évaluer
    - Annotation des corpus
  - Limites
    - Évaluation globale uniquement
- Explorer et définir des méthodes d'évaluation complémentaires

Références

Pour cette présentation :

Evaluation et collection-tests en recherche d'information et catégorisation de textes (Jacques Savoy, Université de Neuchâtel)
EARIA 2012 : http://www.asso-aria.org/documents/earia/Slides/IRTaskCorpus4.pdf

De l'évaluation en traitement automatique des langues (Patrick Paroubek), Habilitation à diriger des recherches (HDR) 2013
http://perso.limsi.fr/pap/hdr_slides_pap.pdf,
http://perso.limsi.fr/pap/hdr_memoir_pap.pdf

Évaluation transparente du traitement des éléments de réponse à une question factuelle, Sarra El Ayari, Thèse de doctorat de l'université Paris Sud, 2009
http://www.linguist.univ-paris-diderot.fr/~sayari/these.php

→ A consulter pour les références dans le domaine de l'évaluation