Évaluation en TAL

Cours de Traitement Automatique des Langues

A.-L. Ligozat

Dernière mise à jour : novembre 2016

Enjeux du cours

- **Enjeux** scientifiques et technologiques de l'évaluation appliquée au TAL
- **Concepts** de l'évaluation
- **Approche concrète** de l'évaluation : tâches, métriques, ressources, protocoles...

- Pointeurs et références pour applications et projets

Plan du cours

- Principes de l'évaluation
- Campagnes d'évaluation
- Trois domaines :
  - analyse syntaxique
  - questions-réponses
  - traduction automatique

Principes

Que signifie évaluer ?

- «Action d'évaluer, d'apprécier la valeur (d'une chose); technique, méthode d'estimation.» (TLFi)
- TAL = démarche **scientifique** + **technologique**
  - important pour estimer le succès d'une recherche
- évaluation de la capacité fonctionnelle (cf. qualité d'un logiciel)

Le paradigme d'évaluation

Évaluation

=

expériences

+

comparaisons

HDR P. Paroubek

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

Niveaux

- évaluation de la **recherche** de base
  - validation d'une nouvelle idée ou mesure du niveau d'innovation
- évaluation **technologique**
  - performance et efficacité d'une technologie pour un problème
- évaluation d'**usage**
  - niveau d'utilisation (usability) d'une technologie pour résoudre un problème réel avec participation de l'utilisateur final
- évaluation d'**impact**
  - conséquences socio-économiques d'une technologie
- évaluation de **programme**
  - atteinte des objectifs finaux d'un programme

Consortium ELSE, Evaluation in Language and Speech Engineering
http://www.limsi.fr/TLP/ELSE/

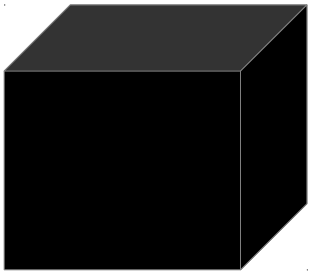
Critères : intrinsèque/extrinsèque

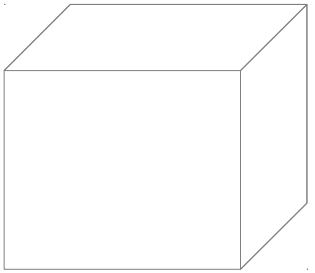
- **intrinsèque** : système évalué par rapport à une référence (gold standard)
- **extrinsèque** ou évaluation *en contexte* : évaluation dans un système complet répondant à une fonction précise pour l'utilisateur

- 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

Boîte noire/transparente

- Génie logiciel
- **Boîte noire** : sorties pour une entrée donnée + éventuellement évaluation de performance, i.e. vitesse, fiabilité, ressources etc.
  - Évaluation **modulaire** : division en sous-tâches
- **Boîte transparente** : évaluation de la structure interne
  - composants du système, ressources linguistiques utilisées, phénomènes linguistiques traités...





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
  - 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)...

    $\Rightarrow$ adjudication éventuelle

Comment évaluer ?

- Capacité fonctionnelle représentée par un ensemble d'attributs
- Métriques pour associer un niveau de qualité à un système pour chaque attribut

- Évaluation individuelle ou collective
  - campagnes d'évaluation comparative

Différents types

- Compétition
  - 1 critère, ordre total, pas d'analyse de performance, pas reproductible
- Validation
  - Plusieurs critères, ordre partiel, seuil d'acceptabilité (performance oui/non), reproductible
- Évaluation
  - Plusieurs critères, ordre partiel, analyse de performance, reproductible

Problématique scientifique

- Définir un champ expérimental commun
  - données
  - représentations
  - mesures
- Pour des problèmes universels
- Et en dégager des résultats reproductibles

HDR P. Paroubek

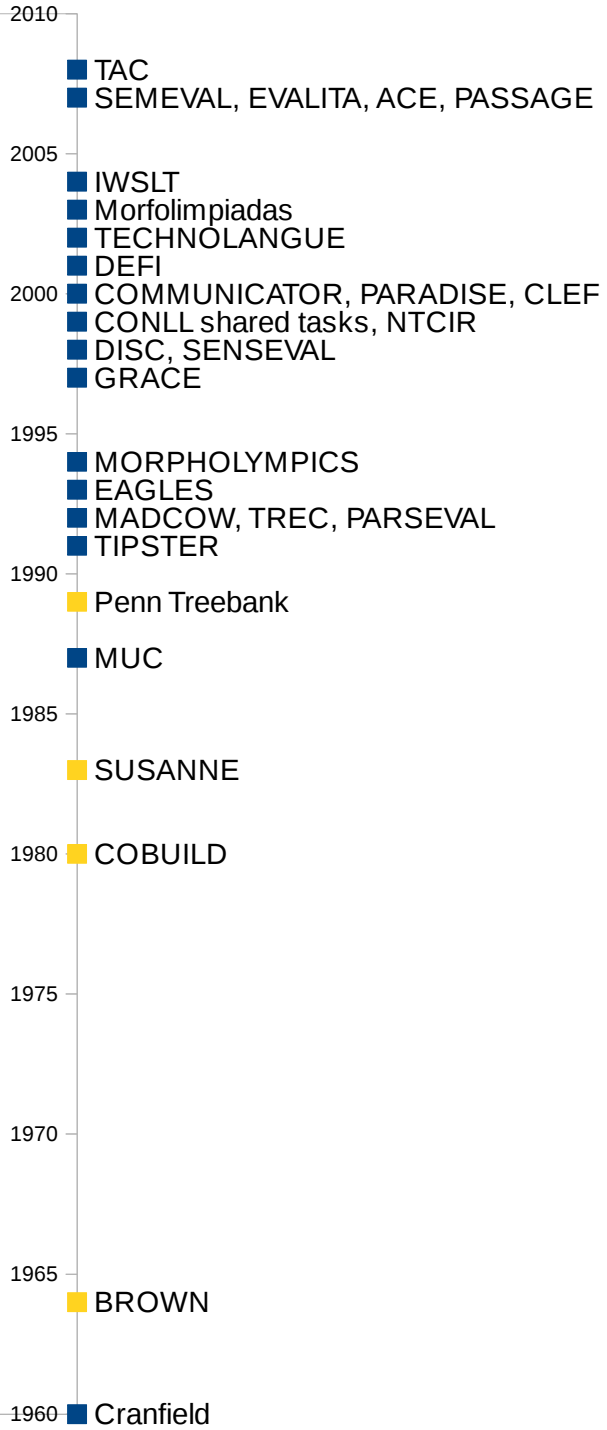
Campagnes d'évaluation

Les débuts

- premières campagnes : tests de Cranfield dans les années 1960, puis rapport ALPAC en 1966
- intérêt accru depuis la fin des années 1990
  - campagnes, participants, tâches => paradigme de l'évaluation
- États-Unis (NIST et DARPA), Japon, Chine, Europe

- États-Unis
  - NIST (National Institute of Standards and Technology) : développement de standards et évaluations
  - DARPA (Defense Advanced Research Projects Agency) : R&D pour défense

Quelques campagnes



Domaines d'évaluation

Composants

- Morphologie
- Terminologie
- Syntaxe
- Entités nommées
- Analyse thématique
- Sémantique
- Anaphore

Fonction principale

- Questions-réponses
- Recherche d'information
- Dialogue
- Extraction d'information

Structure d'une campagne

- Définition de la tâche de contrôle
  - = traitement que les systèmes exécutent au cours de l'évaluation dans les conditions déterminées préalablement
- Métriques
  - Niveau de performance doit être déterminé par algorithme par rapport à une référence (ou étalon ou gold standard) et/ou intervention humaine (attention à reproductibilité alors)
- Ressources
  - annotation manuelle ou semi-automatique
- Cycle de vie
  - 1 à 2 ans

Déroulement d'une campagne

- Phase d'entraînement
  - facultative mais intérêt réel !
  - données d'entraînement représentatives du test
  - protocole communiqué aux participants, ainsi qu'outils
- Phase d'apprentissage
  - ”test à blanc”
- Test
  - Temps imparti pour le traitement des données et fourniture des résultats, mode de fourniture, format prédéfini
  - Résultats doivent être représentatifs, comparables et reproductibles
- Analyse des résultats
  - Atelier de présentation de la campagne
  - Discussions : production de nouvelles ressources, définition de futures métriques et protocoles pour campagnes à venir, identification d'axes de recherche
  - ”kit d'évaluation”

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 gros volumes de données à faible coût ?

HDR P. Paroubek

Kits d'évaluation
(exemples EVALDA)

Campagne	Tâche	Ressources	#documents	#mots
EASy	Analyse syntaxique	Corpus littéraire		150 000
		Corpus oral		8 000
		Corpus de courriels		121 000
		Corpus médical		100 000
		Corpus journalistique et parlementaire		250 000
ESTER	Reconnaissance de la parole	Corpus audio transcrit d'émissions radio-difusées	100 heures	1,1 million de mots
		Corpus audio non transcrit	1 700 heures	-
MEDIA	Dialogue oral	Corpus audio de dialogues enregistrés	1 258 dialogues	
		Corpus de dialogues transcrits	1 258 dialogues	
		Corpus de test annoté en méta-annotations	200 dialogues	438 000 mots

Aspects juridiques et économiques

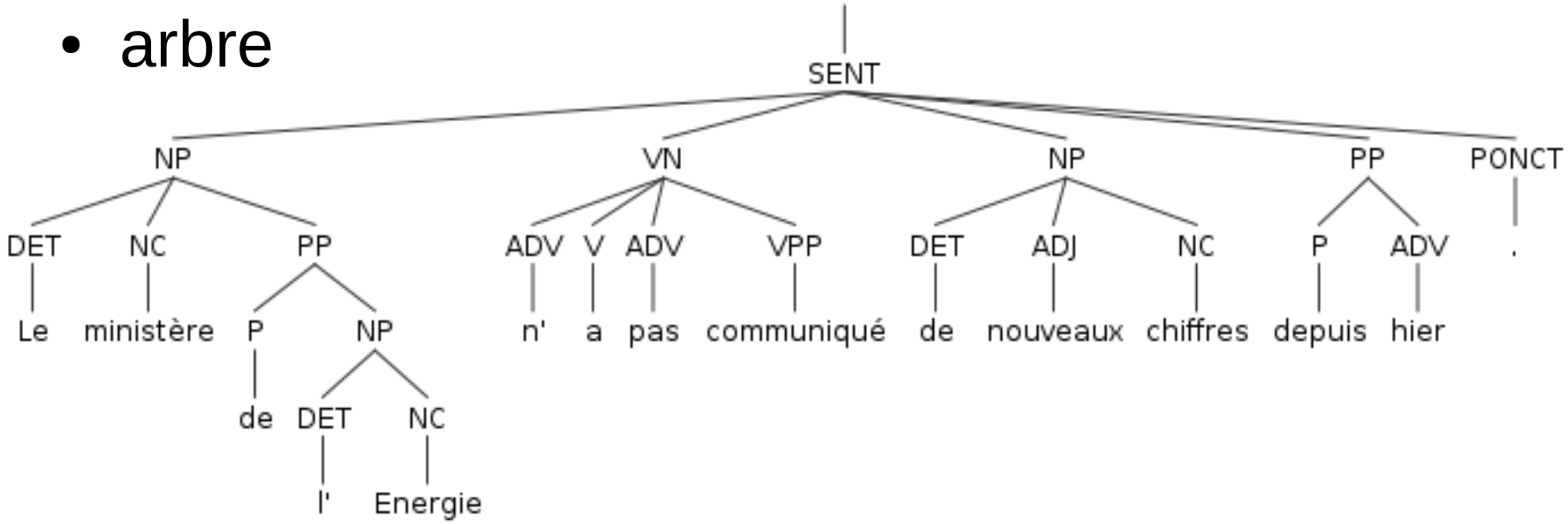
- Aspects juridiques
  - cadre juridique pour les procédures d'évaluation ?
    - UE
  - organisateurs ? (pas de juge et partie)
  - disponibilité des corpus et ressources linguistiques ?
  - exploitation des résultats ?
- Coût d'une évaluation
  - ressources linguistiques pour l'entraînement, de développement et de test
  - évaluation des résultats
  - atelier final
  - ex : traduction automatique dans EVALDA : ~260k€

Analyse syntaxique

Analyse syntaxique

Le ministère de l'Energie n'a pas communiqué de nouveaux chiffres depuis hier.

- Constituants
  - arbre

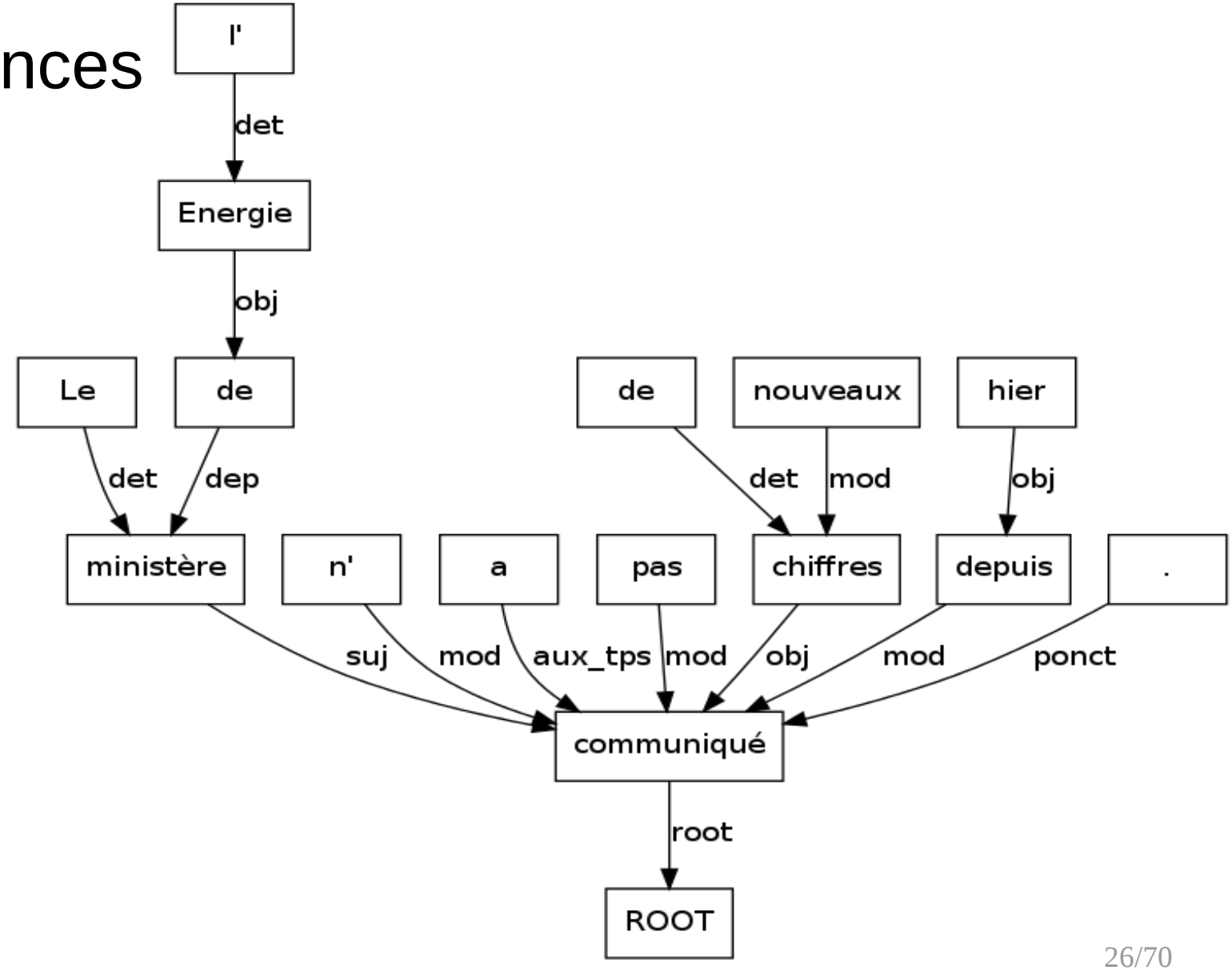


  - parenthésé

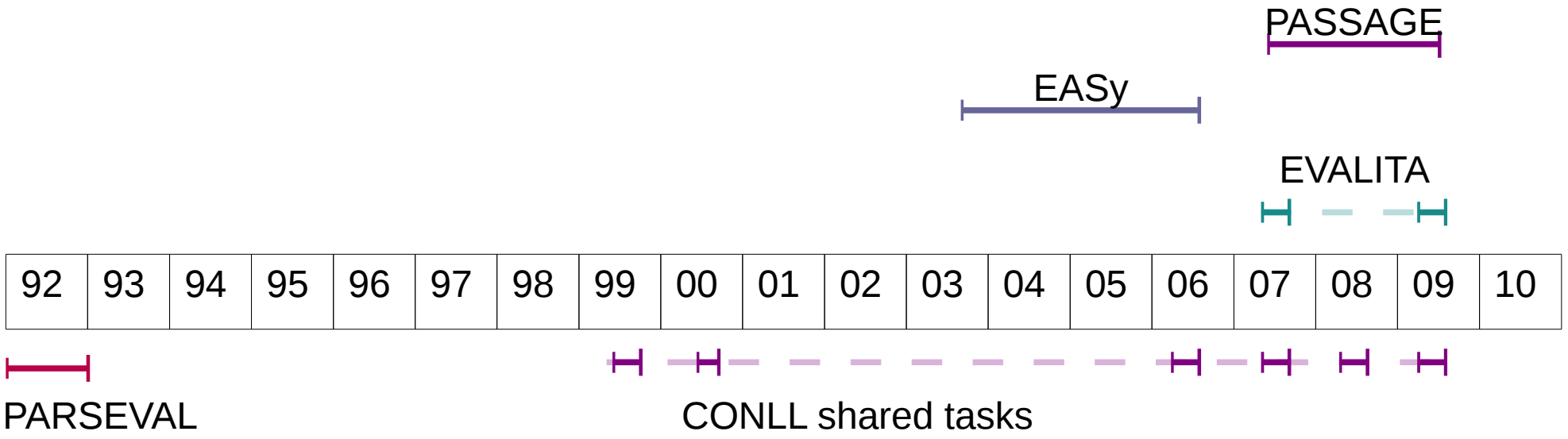
((SENT (NP (DET Le) (NC ministère) (PP (P de) (NP (DET l') (NC Energie)))) (VN (ADV n') (V a) (ADV pas) (VPP communiqué)) (NP (DET de) (ADJ nouveaux) (NC chiffres)) (PP (P depuis) (ADV hier)) (PONCT .)))

Analyse syntaxique

- Dépendances

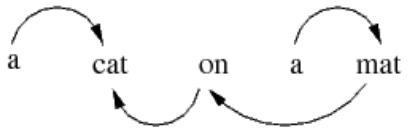


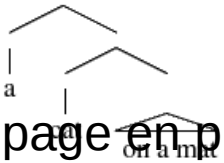
Campagnes d'évaluation

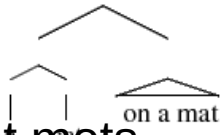


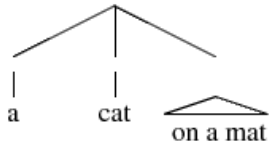
Comparaison d'analyses

- Problème du format unique
  - nombreuses théories syntaxiques...
  - subjectivité
  - exemple : relations sur têtes (mais lesquelles ?) ou constituants ?









- Normalisation du corpus: prédécoupage en phrases et mots
  - problème non résolu !
    - phrases
      - énumérations par exemple
      - encore plus difficile si transcriptions
    - mots
      - mots composés
      - locutions
  - idéalement, géré par systèmes puis réalignement automatique

Annotations EASY

(Paroubek et al., 2008)

- EASy : analyse syntaxique du français
  - guide d'annotation
  - outils d'annotation
  - estimation des erreurs dans l'annotation par un expert
- Constituants minimaux continus non récursifs
  - groupe nominal (« la très grande porte »)
  - groupe prépositionnel (« de la chambre »)
  - noyau verbal (« ne veut »)
  - groupe adjectival (« intacts »)
  - groupe adverbial (« aussi »)
  - groupe verbal introduit par une préposition (« de vraiment bouger »)

Annotations EASY

- Relations
  - sujet-verbe
  - auxiliaire-verbe
  - COD-verbe
  - complément-verbe
  - modifieur-verbe (« Jean *dort profondément* »)
  - complémenteur (« Je pense *qu'*il *viendra* »)
  - attribut-sujet/objet
  - modifieur-nom (« l'*unique fenêtre* »)
  - modifieur-adjectif (« la *très belle* collection »)
  - modifieur-adverbe
  - modifieur-préposition
  - coordination
  - apposition
  - juxtaposition

Essai d'annotation

Le Président a livré les grandes orientations de sa politique scolaire lors d'un discours à la Sorbonne.

(Libération)

Constituants

- groupe nominal (« la très grande porte »)
- groupe prépositionnel (« de la chambre »)
- noyau verbal (« ne veut »)
- groupe adjectival (« intacts »)
- groupe adverbial (« aussi »)
- groupe verbal introduit par une préposition (« de vraiment bouger »)

Relations

- sujet-verbe
- auxiliaire-verbe
- COD-verbe
- complément-verbe
- modifieur-verbe (« Jean *dort profondément* »)
- complémenteur (« Je pense *qu'*il *viendra* »)
- attribut-sujet/objet
- modifieur-nom (« l'*unique fenêtre* »)
- modifieur-adjectif (« la *très belle* collection »)
- modifieur-adverbe
- modifieur-préposition
- coordination
- apposition
- juxtaposition

Corpus arborés (Treebanks)

- Corpus dans lequel chaque phrase est analysée
- Penn Treebank pour l'anglais
- French Treebank pour le français (Abeillé et al., 2003)
  - constituants et fonctions
  - corpus journalistique
  - 1 million de mots

Métriques

- précision, rappel, f-mesure (RI) (Rijsbergen, 1979)

$$p = \frac{R \cap H}{H} \qquad r = \frac{R \cap H}{R} \qquad f = \frac{1}{\frac{\alpha}{p} + \frac{1-\alpha}{r}} \overset{\alpha=0,5}{=} \frac{2 \times p \times r}{p + r}$$

R = référence
H = hypothèse (système)

- valable pour constituants et dépendances (étiquetés ou non)

Exemple (CoNLL 99)

Référence :

Les raffineries votent la reprise.
GN GV GN

Hypothèse :

Les raffineries votent la reprise.
GN GV GA

p_{GN} = 1/1, r_{GN} = 1/2, f_{GN} = 2/3

Exemple pour dépendances

- Kim promised Alex to bring some wine.

Référence :
(Kim, NNP, promised, subj)
(promised, VBD)
(Alex, NNP, promised, **obj1**)
(to, TO, bring)
(bring, VB, promised, obj2)
(some, DT, wine)
(wine, NN, bring, obj1)

Hypothèse :
(Kim, NNP, promised, subj)
(promised, VBD)
(Alex, NNP, promised, **subj**)
(to, TO, bring)
(bring, VB, promised, obj1)
(some, DT, wine)
(wine, NN, bring, obj1)

Sans étiquettes :
précision = 5/6, rappel = 5/6
Avec étiquettes :
précision = 4/6, rappel = 4/6

Extensions

- Cross brackets (PARSEVAL) (Black et al., 1991)
  - Nombre de constituants de l'hypothèse qui croisent un constituant de la référence
    - ((A B) C) vs. (A (B C))

Exemple (Lin 95)

Référence : (They ((came) yesterday))

Hypothèse : ((They (came)) (yesterday))

- p = 2/4, r = 2/3
- 1 pair of cross brackets : [0,1] et [1,2]

(Sampson et Babarczy, 2003)

    - pénalise fortement certaines erreurs de rattachement
- Leaf-ancestor metric
  - Compare les ancêtres des mots avec une distance d'édition

Quelques résultats

- Chunking à CoNLL 2000
  - meilleurs systèmes à 94% de précision et rappel
- Dépendances et rôle sémantique à CoNLL 2009
  - meilleurs systèmes entre 85 et 90% de précision et rappel pour l'anglais
- EASy
  - meilleurs systèmes à ~90% de précision et rappel pour les constituants
  - ~60% pour les relations

Questions-réponses

Système de questions-réponses

- Recherche d'information précise
  - entrée : question en langage naturel
  - sortie : réponse précise

When did Alaska become a state ?

When did Alaska become a state ?

When did Alaska become a state ?

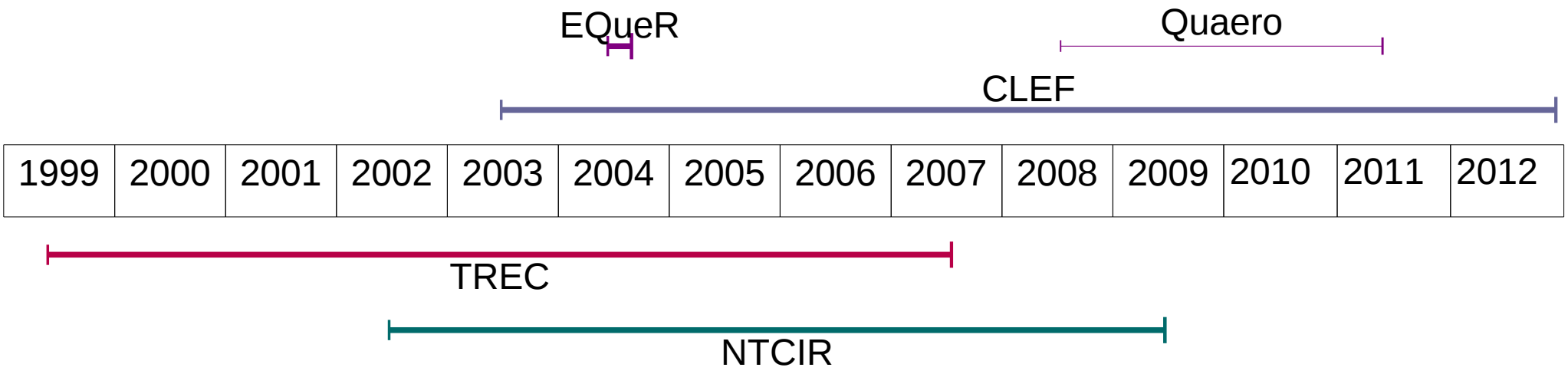
MR

QR

- ALASKA.com|FAQ:How can I become a state park volunteer ?
… How can I become a state park volunteer ?...
- Alaska Elections – State Division of Elections Home Page
- Alaska State Legislature Homepage…

in 1959

Évolution des campagnes



Caractéristiques des campagnes

Campagne	Organisation	Langues traitées	Domaines	Types de questions	Corpus
TREC	NIST (USA)	anglais	général	factuelles, définition, séries	journalistique, blogs
CLEF	divers (Europe)	nombreuses dont anglais, français, italien, espagnol... + translingue	général	factuelles, définition, temporelles	journalistique, wikipédia, transcriptions (manuelles et automatiques), textes parlementaires
NTCIR	NII (Japon)	japonais, chinois + translingue	général	factuelles, définition, listes complexes (relations...)	journalistique, web (restreint)
EQueR	LIMSI/ELDA (France)	français	général, médical	factuelles, définition, liste, oui/non	journalistique, sénat, articles scientifiques, recommandations de bonne pratique médicale
Quaero	Synapse (France)	anglais, français	général	factuelles, définition, oui/non	web

Types de questions «classiques»

- factuelles
  - entités nommées : Combien l'Islande compte-t-elle d'habitants ?
- définition
  - personnes : Qui était Federico García Lorca ?
  - groupes : Qui sont les BB Brunes ?
  - autres : Qu'est-ce que Bollywood ? Qu'est-ce qu'une galère ?

Types de questions étendus

- listes
  - Citez des boxeuses professionnelles.
- temporelles (CLEF 2005)
  - Quel journal a été fondé à Kiev en 1994 ? Qui était le président des États-Unis entre 1976 et 1980 ? Quelle maladie de nombreux soldats américains ont-ils contractée après la guerre du golfe ?
- oui/non
  - La carte d'identité existe-t-elle au Royaume Uni ?
- à thème (TREC 2004)
  - Qui a fondé les Black Panthers ? Quand ont-elles été fondées ? Où ? Qui a été membre de cette organisation ? Autres
- relations
  - Quelle est la relation entre Nicolas Cage et Francis Ford Coppola ?

Réponses

- passage de texte contenant la réponse de 250 caractères (1er TREC)
- puis réponse précise
  - justifiée par passage

Quelle monnaie utilise-t-on en Argentine ?

Le **nouveau peso argentin**, dont le code ISO est ARS, est la monnaie officielle de l'Argentine.

- QCM (QA4MRE)

- quelques centaines de questions par évaluation
- évaluations translingues
  - question dans langue source
  - documents (et réponse) dans langue cible différente

Critères de jugement d'une réponse

- **pertinence**
  - la réponse doit répondre à la question
  - «How tall is the statue of Liberty?» ⇒ réponses donnant tailles des copies de la statue considérées comme fausses
- **précision**
  - la réponse doit se situer au bon niveau de granularité
  - «Where was Harry Truman born?» ⇒ réponses «Lamar, Missouri» et «Missouri», correctes, mais pas «USA»
- **concision**
  - la réponse ne doit pas contenir d'information inutile
  - «What river in the US is known as the Big Muddy? » ⇒ «Big Muddy, the Mississippi is the longest» inexacte

Critères de jugement d'une réponse

- **complétude**
  - la réponse ne doit pas être partielle
  - «500» n'est pas une bonne réponse si« $500» attendue
- **simplicité**
  - l'utilisateur doit pouvoir lire la réponse facilement
- **justification**
  - passage justificatif fourni avec réponse
- **contexte**
  - passage doit permettre de déterminer le contexte de validité de la réponse
  - Who is the French president ? ⇒ date du document importante

(Ligozat, 2006)

Corpus

- journalistique
  - Le Monde, Financial Times, LA Times, China Times...
- web
  - Wikipédia, blogs, communautaire...
  - contient information semi-structurée comme tableaux, titres...
    - peuvent contenir informations intéressantes mais difficiles à exploiter
  - documents de spam
  - reconnaissance de la langue du document
  - problèmes de conversion de documents PDF, DOC...

(Tannier & Moriceau, 2010)

- taille : plusieurs centaines de Mb à plusieurs dizaines de Gb

Évaluation : jugements

- manuels !
- 2 jugements : passage et réponse courte
  - réponse courte
    - correcte (comprend NIL i.e. pas de réponse)
    - inexacte : bonne réponse mais pas suffisamment précise
    - incorrecte : mauvaise réponse
    - non justifiée : bonne réponse mais pas justifiée par le passage
  - passage
    - correct
    - incorrect

Évaluation : métriques

(Voorhees, 2000)

- standard : MRR (Mean Reciprocal Rank)
  - moyenne des inverses des rangs des premières réponses correctes

$$MRR = \frac{1}{\#questions} \sum_{i=1}^{\#questions} \frac{1}{answer_i rank}$$

Question	Réponse 1	Réponse 2	Réponse 3	Reciprocal rank
Quelle est la durée d'un mandat de député ?	3	**5**	7	1/2
Quel est l'âge minimum pour être député ?	**23 ans**	27 ans	25 ans	1
Qui est le président de l'Assemblée ?	Jean-Louis Debré	Patrick Ollier	**Bernard Accoyer**	1/3
MRR : (1/3 + 1/2 + 1) / 3 = 11/18 ≃ 0.6				

Évaluation : métriques

- NIAP (Non Interpolated Average Precision) pour les questions listes
  - but : favoriser les réponses les plus complètes

$$NIAP = \frac{\sum \frac{i}{r_i}}{T}$$

où r_i est le rank du document i
et T est le nombre de documents pertinents dans la collection

Évaluation : discussion

- complétude de la réponse
  - jour et mois dans la réponse, année dans le titre
- granularité de la réponse
  - Où se trouve la Joconde ?
  - au musée du Louvre, à Paris, en France...
- représentativité de la réponse
  - Qu'est-ce que l'Atlantis ?

- Dans la mythologie grecque, ***Atlantis*** est la prononciation grecque du nom akkadien Atrahasis, signifiant *le sage des sages* ou des racines *ḫaṭṭu ḫasīsu* "sceptre de l'ingéniosité", et désignant le Dieu/Roi titanide à l'origine de toutes nos civilisations.

Sommaire [masquer]

Littérature et cinéma [modifier]

- **Atlantis** est un roman de David Gibbins publié en 2005 ;
- **Atlantis** est une revue fondée en 1927, dirigée aujourd'hui par un des disciples de Paul Le Cour, Jacques d'Arès. C'est une des revues notables de l'ésotérisme contemporain ;
- **Atlantis** est un film réalisé en 1930 par Ewald-André Dupont & Jean Kemm ;
- **Atlantis** est un film documentaire réalisé par Luc Besson en 1991 ;
- **Atlantis, l'empire perdu** est le titre québécois de Atlantide, l'empire perdu (*Atlantis, the Lost Empire*), un film des Studios Disney sorti en 2001 ;
- **Stargate Atlantis** est une série télévisée américano-canadienne de science-fiction, dérivée de la série *Stargate SG-1*.

Musique [modifier]

- **Atlantis** est le premier album du groupe de musique Lunatica.

Jeux vidéo [modifier]

- ***Atlantis*** est un des premiers jeux d'aventure graphique français sur Oric, Amstrad CPC, MO5 et ZX Spectrum édité par Cobrasoft en 1985 ;
- ***Atlantis, secrets d'un monde oublié*** est le premier jeu de la série de jeux vidéo *Atlantis* développé par Cryo Interactive ;

Lieux, réels ou fictifs [modifier]

- **Atlantis** est un lieu de la série télévisée Stargate Atlantis ;
- **Atlantis** est la plus grande zone commerciale de l'ouest de la France, située dans l'agglomération nantaise ;
- **Le Défi d'Atlantis** était le nom de l'une des attractions et du film que celle-ci projetait au parc Futuroscope, à Poitiers, en France.
- Dans le monde des bandes-dessinées Marvel, **Atlantis** est la cité sous-marine du peuple atlante gouvernée par Namor.
- **Atlantis** est le nom de l'île mythique des mages dans le jeu de rôle Mage : l'Éveil

Vaisseaux, réels ou fictifs [modifier]

- **Atlantis** est une navette spatiale américaine.
- **Atlantis** est un navire de guerre allemand, camouflé en navire de commerce pendant la seconde guerre mondiale.
- L' **Atlantis** est un trois-mâts barquentine des Pays-Bas.
- L' Atlantis a été un paquebot de croisière britannique précédemment appelé Andes
- **Atlantis** est un fabricant de yachts.
- **Atlantis** est le nom du vaisseau d'Albator dans la version française des dessins animés japonais *Albator 78* et *Albator 84*. (Dans la version originale il s'appelle en fait *Arcadia*).
- **Atlantis** est une cité-vaisseau fictive où se déroule la trame principale de la série Stargate Atlantis.

Informatique [modifier]

- Atlantis est un logiciel propriétaire de traitement de texte développé par la société Rising Sun Solutions.

Quelques résultats

- ResPubliQA (CLEF 2010)
  - documents parlementaires européens
  - questions factuelles, définition, raison/but, procédure, opinion, autres
  - langues : allemand, anglais, espagnol, français, italien, portugais, roumain (peu de soumissions interlingues)
  - meilleur système à 70% de bonnes réponses sur l'anglais
- Quaero
  - corpus web
  - questions factuelles, listes, définitions, oui/non...
  - langues : français et anglais
  - meilleurs systèmes à MRR ~ 0.4 en anglais et français

Traduction automatique

Traduction anglais → français : exemple

Bataille De la Grande-Bretagne

J’ai toujours été un ventilateur de ce film et avais impatiemment attendu un dégagement de DVD pendant un certain temps. Quand j’ai découvert que MGM libéraient ce film sur DVD, j’ai été enchanté ai suivi d’une certaine irritation quand j’ai découvert le RRP 19,99, £ car le film était dehors aux USA pour $8. Encore un autre disque d'édition spéciale deuxièmes des morceaux avec augmenté vers le haut de l'étiquette des prix.

Je fais des excuses par ceci à MGM pour de telles pensées d’unkind, parce que cette édition est superbe. La version des USA est dans mono, où ce dégagement a 5,1 et DTS (il est excellent) et pour des ventilateurs de William Walton (ils ont seulement employé environ 5 minutes de ses points dans le film original, Ron Goodwin assurant le repos), la bande sonore avec les pleins points de Walton comme alternative. L’image a été également reconstituée et a une pleine image anamorphic de 2.35 :1.

Traduction anglais → français : exemple

Bataille De la Grande-Bretagne

J'ai toujours été un ventilateur (*fan*) de ce film et avais impatiemment attendu un dégagement (*release*) de DVD pendant un certain temps. Quand j'ai découvert que MGM libéraient (*were releasing*) ce film sur DVD, j'ai été enchanté ai suivi d'une certaine irritation quand j'ai découvert le RRP 19,99, £ (*the RRP (£19.99)*) car le film était dehors aux USA (*was out in the US*) pour $8. Encore un autre disque d'édition spéciale deuxièmes des morceaux avec augmenté vers le haut de l'étiquette des prix (*with a hiked up price tag*).

Je fais des excuses par ceci à MGM pour de telles pensées d'unkind, parce que cette édition est superbe. La version des USA est dans mono, où ce dégagement a 5,1 et DTS (il est excellent) et pour des ventilateurs de William Walton (ils ont seulement employé environ 5 minutes de ses points (*score*) dans le film original, Ron Goodwin assurant le repos) (*Ron Goodwin supplying the rest*), la bande sonore avec les pleins points de Walton comme alternative. L'image a été également reconstituée et a une pleine image anamorphic de 2.35 :1.

Battle of Britain = La Bataille d'Angleterre
RRP : prix recommandé
score : partition

Autre exemple

Claire, si ça ne t'ennuie pas, on peut se tutoyer. Après tout, quand on s'est connus, tu avais 15 ans. C'était l'hiver 1995-1996 et tu t'appelais Angela Chase. La série My so-called life avait déjà pris fin aux Etats-Unis mais, avec l'habituel décalage, on se retrouvait chaque semaine, via Canal Jimmy, au Liberty High School, un lycée de la banlieue de Pittsburgh. Même quand on n'était pas dans le cœur de cible, on mordait à Angela, 15 ans. Une série peuplée de teen-agers pas vraiment lobotomisés par le fun. Avec un bellâtre illettré, un chicano gay, une peste mi-punk et un blême échalas transi, tel Charlie Brown, pour sa « petite fille rousse ». Tu prêtais ton visage idéalement boudeur à cette ado introvertie qui tenait la chronique de sa « prétendue vie ». Seule et nous mettant tous dans la confidence, tu observais tes congénères.

Claire, if it does not bother you, you can tu. After all, when it was known you had 15 years. It was the winter of 1995-1996 and you called Angela Chase. Series My so-called life had ended in the United States, but with the usual lag, we found each week via Canal Jimmy, at Liberty High School, a high school in suburban Pittsburgh. Even when we were not in the heart of target, biting Angela, 15 years. A series of teen-agers populated not by brainwashed really fun. A fop illiterate, a gay Chicano, a plague mid-punk and pale stakes transition, as Charlie Brown, for his "little redheaded girl." You lent your face perfectly to this sulky teenager who was introverted chronicling his "alleged life." Alone and putting us all in the know, you watched your peers.

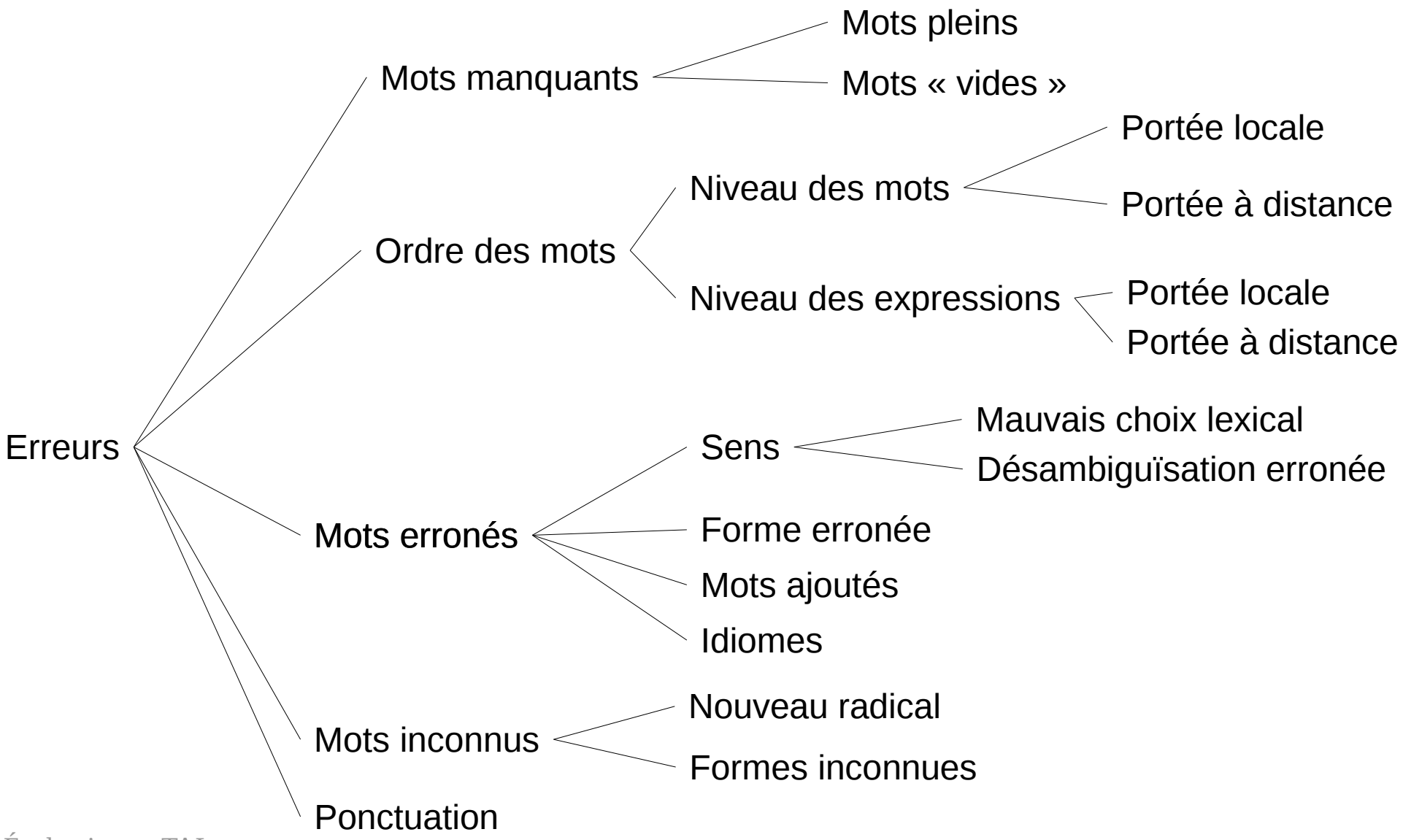
Article de Télérama, octobre 2012
Avec “Homeland”, Claire Danes quitte enfin “Angela, 15 ans”
Traduction par Google Translate

Évaluation des systèmes de traduction automatique

- Critères d'évaluation possibles :
  - attributs intrinsèques comme fidélité au texte source et caractère naturel du texte cible
  - évaluation par la tâche
  - coûts et bénéfices
- Besoin d'évaluation rapide
  - performances relatives de systèmes
  - ou de versions d'un même système
- Critères d'évaluation difficiles même pour un humain
- Métriques proposées sujettes à critiques mais jugement humain coûte cher
- Évaluation automatique doit :
  - être rapide
  - s'appliquer à plusieurs langues
  - corréler au mieux avec le jugement humain

Typologie des erreurs de traduction

(Vilar et al., 2006)



BLEU (BiLingual Evaluation Understudy*)

(Papineni, 2002)

- Hypothèse : plus un texte traduit automatiquement ressemble à un texte produit par un traducteur professionnel humain, meilleure est sa qualité
  - permettre une évaluation rapide des systèmes de TA
  - comparer une traduction avec une ou plusieurs traductions de référence sur la base de groupes de mots en commun
  - fondé sur :
    - une mesure de proximité de traduction inspirée de la mesure du taux d'erreur sur les mots (*word error rate* en parole)
    - un corpus de textes traduits par des traducteurs professionnels

* suppléant pour les évaluations bilingues

Exemples

- Référence 1: It is a guide to action that ensures that the military will forever heed Party commands.
- Référence 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Référence 3: It is the practical guide for the army always to heed the directions of the party.

- Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.
- Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

• Candidate 1 clairement meilleur que candidate 2
  - plus de n-grams communs avec les trois références

Calcul de scores BLEU

Moyenne pondérée des précisions n-gram

$$score = BP * \exp\left(\sum_{n=1}^{N} w_n \log\left(\underbrace{\frac{|ngrams_{trad} \cap ngrams_{ref}|}{|ngrams_{trad}|}}_{p_n}\right)\right)$$

avec $w_n = \frac{1}{N}$ (poids uniformes)

$N = 4$

$$BP = \exp\left(1 - \frac{|ref|}{|trad|}\right) si\ |trad| \leqslant |ref|, 1\ sinon$$

(*brevity penalty* : si la traduction est trop courte, score baisse)

Exemple de calcul de score (Déchelotte, 2007)

I am feeling good

p1 = 4/4=1
p2 = 2/3
p3 = 1/2
p4 = 0

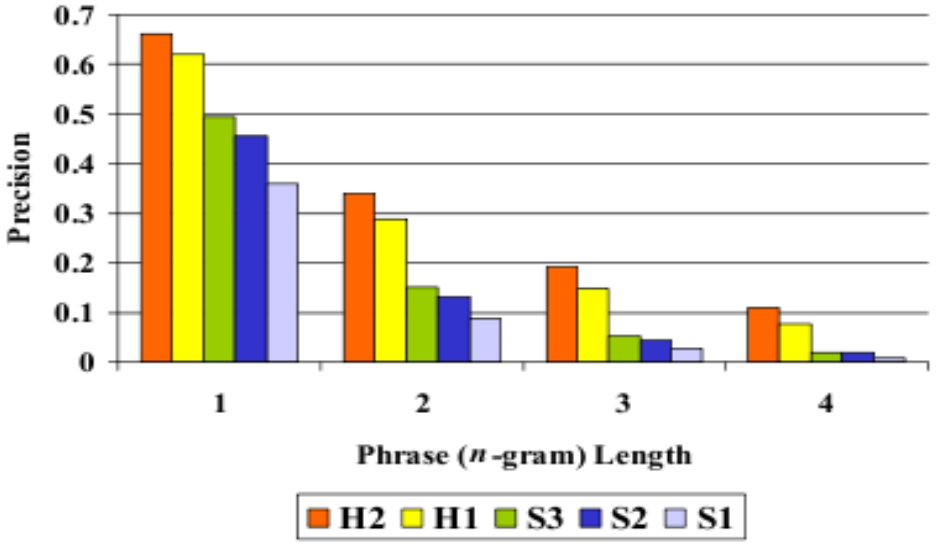
Réf1 : I am happy

Réf2 : I am feeling very good

Points forts de BLEU

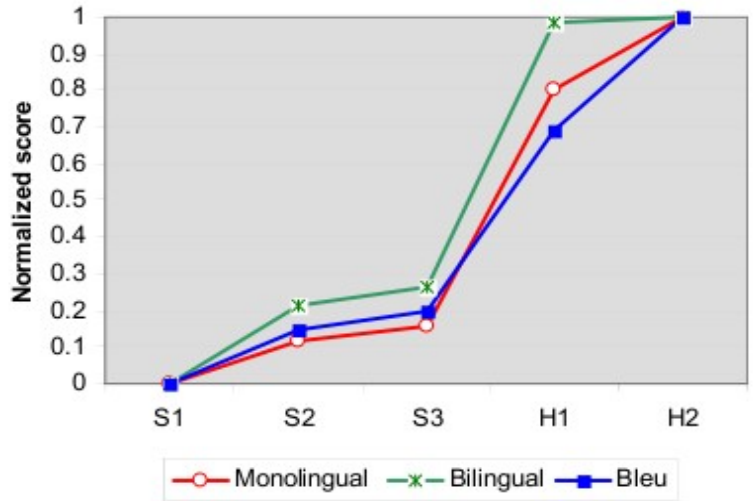
- Évaluation par BLEU
  - de trois systèmes (S1, S2, S3)
  - d'un humain n'ayant ni la langue source ni la langue cible comme langue maternelle (H1)
  - et d'un traducteur anglophone (H2) vers l'anglais

  ⇒ H2 > H1 > S3, S2, S1



- Comparaison avec jugements humains
  - monolingue = locuteurs anglais
  - bilingue = locuteurs chinois vivant aux US

  ⇒ forte corrélation avec évaluation humaine



Essai d'évaluation de traductions

To find out who shaded the truth most, TIME asked each campaign for a list of its rival's worst deceptions. After examining those claims and consulting independent fact-checking websites, we selected some of the most prominent falsehoods and prevarications of the 2012 campaign*—at least so far. Compared with the Obama campaign's, the Romney operation's misstatements are frequently more brazen. But sometimes the most effective lie is the one that is closest to the truth, and Obama's team has often outdone Romney's in the dark art of subtle distortion. On both sides, the dishonesty is "about as bad as I've seen," says veteran journalist Brooks Jackson, director of FactCheck.org.

Pour découvrir qui ont ombragé la vérité plus, le TEMPS a demandé chaque campagne une liste des plus mauvaises duperies de son rival. Après examen de ces réclamations et consultation des sites Web de fait-vérification indépendants, nous avons choisi certains des mensonges et des tergiversations les plus importants du campaign*-at 2012 mineurs jusqu'ici. Comparé à la campagne d'Obama, les rapports inexacts de l'opération de Romney sont fréquemment plus d'airain. Mais parfois le mensonge le plus efficace est celui qui est le plus proche de la vérité, et l'équipe d'Obama a souvent surpassé Romney dans l'art foncé de la déformation subtile. Des deux côtés, la malhonnêteté est « environ aussi mauvaise que j'ai vu, » dit le journaliste Brooks Jackson, directeur de vétéran de FactCheck.org.

systran

Essai d'évaluation de traductions

To find out who shaded the truth most, TIME asked each campaign for a list of its rival's worst deceptions. After examining those claims and consulting independent fact-checking websites, we selected some of the most prominent falsehoods and prevarications of the 2012 campaign*—at least so far. Compared with the Obama campaign's, the Romney operation's misstatements are frequently more brazen. But sometimes the most effective lie is the one that is closest to the truth, and Obama's team has often outdone Romney's in the dark art of subtle distortion. On both sides, the dishonesty is "about as bad as I've seen," says veteran journalist Brooks Jackson, director of FactCheck.org.

Pour savoir qui ombrageaient la vérité, la plupart TEMPS demandé à chaque campagne pour une liste des pires tromperies de son rival. Après examen de ces demandes et de consultation indépendants vérification des faits sites Internet, nous avons sélectionné quelques-uns des mensonges les plus importants et prévarications de la campagne de 2012 au-* à ce jour. Par rapport à la campagne d'Obama, les anomalies du fonctionnement de Romney sont souvent plus éhontée. Mais parfois, le mensonge le plus efficace est celui qui est le plus proche de la vérité, et l'équipe d'Obama a souvent surpassé Romney dans l'art sombre de la distorsion subtile. Des deux côtés, la malhonnêteté est «à peu près aussi mauvais que j'ai vu», affirme le journaliste vétéran Brooks Jackson, directeur de FactCheck.org.

google

Essai d'évaluation de traductions

To find out who shaded the truth most, TIME asked each campaign for a list of its rival's worst deceptions. After examining those claims and consulting independent fact-checking websites, we selected some of the most prominent falsehoods and prevarications of the 2012 campaign*—at least so far. Compared with the Obama campaign's, the Romney operation's misstatements are frequently more brazen. But sometimes the most effective lie is the one that is closest to the truth, and Obama's team has often outdone Romney's in the dark art of subtle distortion. On both sides, the dishonesty is "about as bad as I've seen," says veteran journalist Brooks Jackson, director of FactCheck.org.

Pour découvrir qui a hachuré la vérité le plus, le TEMPS a demandé à chaque campagne une liste des pires tromperies de son rival. Après l'examen de ces réclamations et conseil de sites Web vérifiant fait indépendants, nous avons choisi certains des mensonges les plus en vue et les tergiversations du 2012 campaign*- at le moins jusqu'ici. Comparé avec la campagne d'Obama, les déclarations erronées de l'opération Romney sont fréquemment plus sans gêne(de cuivre). Mais parfois le mensonge le plus effectif(efficace) est celui qui est le plus proche à la vérité et l'équipe d'Obama surpassait souvent Romney dans l'art sombre d'altération subtile. Des deux côtés, la malhonnêteté est "d'aussi mal que j'ai vu," dit que le journaliste expérimenté Tolère Jackson, le directeur de FactCheck.org.

reverso

Limitations de BLEU

- Évaluation sur phrases individuelles : pas vraiment de sens
- Nombre de traductions de référence nécessaires
  - nécessité de prendre en compte plusieurs formulations possibles d'une même traduction
  - expériences avec une seule traduction de référence choisie au hasard parmi 4 donnent classements comparables : gros corpus de test avec une seule traduction par phrase suffisant
- Meilleur choix de traduction ne résulte pas nécessairement en l'amélioration du score
- Conclusions
  - besoins d'autres métriques automatiques d'évaluation prenant mieux en compte les variations (beaucoup ont été proposées)
  - besoins d'évaluations manuelles

Conclusion

- Évaluation primordiale en TAL
  - permet la comparaison de méthodes
- Choix
  - métriques
  - jeux de données

Références

- Abeillé, A., Clément, L., & Kinyon, A. (2003). Building a French treebank. Dans A. Abeillé (Éd.), Treebanks (p. 165–188). Kluwer.
- Ayache, C., Grau, B., & Vilnat, A. (2006). EQueR: the French Evaluation campaign of Question-Answering Systems. Dans Proceedings of LREC.
- Black, E., Abney, S., Flickenger, S., Gdaniec, C., Grishman, C., Harrison, P., Hindle, D., et al. (1991). A Procedure for quantitatively comparing the syntactic coverage of English grammars. Dans Proceedings of the workshop on Speech and Natural Language, Human Language Technology Conference. Association for Computational Linguistics. Retrouvé de http://acl.ldc.upenn.edu/H/H91/H91-1060.pdf
- Candito, M., Nivre, J., Denis, P., & Anguiano, E. H. (2010). Benchmarking of Statistical Dependency Parsers for French. Dans Proceedings of COLING'2010 (poster session). Retrouvé de http://alpage.inria.fr/statgram/frdep/Publications/frdepcompar-Coling10-final.pdf
- Chaudiron, S., & Choukri, K. (Éd.). (2008). L'évaluation des technologies de traitement de la langue, les campagnes Technolangues. Hermes.
- Hartley, A., & Popescu-Belis, A. (2004). Évaluation des systèmes de traduction automatique. Dans S. Chaudiron (Éd.), Évaluation des systèmes de traitement de l'information (p. 311–335). Hermes.
- Ligozat, A. (2006). Exploitation et fusion de connaissances locales pour la recherche d'informations précises. Université Paris-Sud 11. Retrouvé de http://www.limsi.fr/Individu/annlor/docs/manuscrit.pdf
- Lin, D. (1995). A Dependency-based Method for Evaluating Broad-Coverage Parsers. Dans Proceedings of IJCAI-95.
- Max, A. (2008). Introduction à la traduction automatique. Cours Université Paris-Sud 11, .
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. Dans Proceedings of the 40th annual meeting on association for computational linguistics (p. 311–318).

Références

- Paroubek, P., Robba, I., & Vilnat, A. (2008). EASY : la campagne d'évaluation des analyseurs syntaxiques. Dans L'évaluation des technologies de traitement de la langue, les campagnes Technolangues (p. 117-140). Hermes.
- Paroubek, P., Chaudiron, S., & Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. TAL, 48(1), 7–31.
- Penas, A., Forner, P., Rodrigo, A., Sutcliffe, R., Forascuand, C., & Mota, C. (2010). Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. Dans Working Notes for the CLEF 2010 Workshop (p. 20-23).
- Rijsbergen, C. J. V. (1979). Information Retrieval (second.). Newton, MA, USA: Butterworth-Heinemann.
- Sampson, G., & Babarczy, A. (2003). A test of the leaf-ancestor metric for parse accuracy. Natural Language Engineering, 9, 365–380.
- Tannier, X., & Moriceau, V. (2010). FIDJI: Web Question-Answering at Quaero 2009. Dans Proceedings of LREC.
- Tou, N. H., Yong, L. C., & King, F. S. (1999). A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. Dans Proceedings of the ACL SIGLEX Workshop on Standarizing Lexical Resources. Retrouvé de http://acl.eldoc.ub.rug.nl/mirror/W/W99/W99-0502.pdf
- Tou, N. H., Lim, C. Y., & Foo, S. K. (1999). A Case Study On Inter-Annotator Agreement For Word Sense Disambiguation. Dans SIGLEX Workshop On Standardizing Lexical Resources.
- Vilar, D., Xu, J., Fernando, D. L., & Ney, H. (2006). Error analysis of statistical machine translation output. Dans Proceedings of the Fifth Int. Conf. on Language Resources and Evaluation (LREC).
- Voorhees, E. M. (2000). Overview of the TREC-9 Question Answering Track. Dans Proceedings of TREC-9 (p. 71–80).