Évaluation humaine ou automatique ?

l'exemple de la Traduction Automatique

Hervé Blanchon
LIG-GETALP
herve.blanchon@imag.fr

Matinée "évaluation", 10 avril 2008

Évaluation des systèmes de TA

- Préoccupation ancienne
  - ALPAC 1996 *(USA)* : travaux de recherche
    - évaluation de l'intelligibilité & de la fidélité
    - *bases de l'évaluation subjective (humaine)*
  - JEIDA 1989-1992 *(Japon)* : systèmes commerciaux
    - premier effort de formalisation
      - ≠ critères ; ≠ échelles de valeur
    - évaluation économique par les utilisateurs
    - évaluation technique par les utilisateurs & les développeurs

Évaluation des systèmes de TA

- Nouvel effort de formalisation (1999 -)
  - FEMTI (Framework for the Evaluation of MT ; UE & NSF)
    - contexte d'usage
    - caractéristiques de qualité des systèmes

2 System characteristics to be evaluated

- 2.1 System internal characteristics
  - 2.1.1 MT system-specific characteristics
  - 2.1.2 Translation process models
  - 2.1.3 Linguistic resources and utilities
  - 2.1.4 Characteristics of process flow
- 2.2 System external characteristics
  - **2.2.1 Functionality**
  - 2.2.2 Reliability
  - 2.2.3 Usability
  - 2.2.4 Efficiency
  - 2.2.5 Maintainability

2.2.1 Functionality

- 2.2.1.1 Suitability
  - 2.2.1.1.1 Target-language only
    - 2.2.1.1.1.1 Readability (or: fluency, intelligibility, clarity)
    - 2.2.1.1.1.2 Comprehensibility
    - 2.2.1.1.1.3 Coherence
    - 2.2.1.1.1.4 Cohesion
  - 2.2.1.1.2 Cross-language / contrastive
    - 2.2.1.1.2.1 Coverage of corpus-specific phenomena
    - 2.2.1.1.2.2 Style
- 2.2.1.2 Accuracy
  - 2.2.1.2.1 Fidelity
  - 2.2.1.2.2 Consistency
  - 2.2.1.2.3 Terminology
- 2.2.1.3 Wellformedness
  - 2.2.1.3.1 Punctuation
  - 2.2.1.3.2 Lexis / lexical choice
  - 2.2.1.3.3 Grammar / syntax
  - 2.2.1.3.4 Morphology
- 2.2.1.4 Interoperability
- 2.2.1.5 Compliance
- 2.2.1.6 Security

Évaluation des systèmes de TA

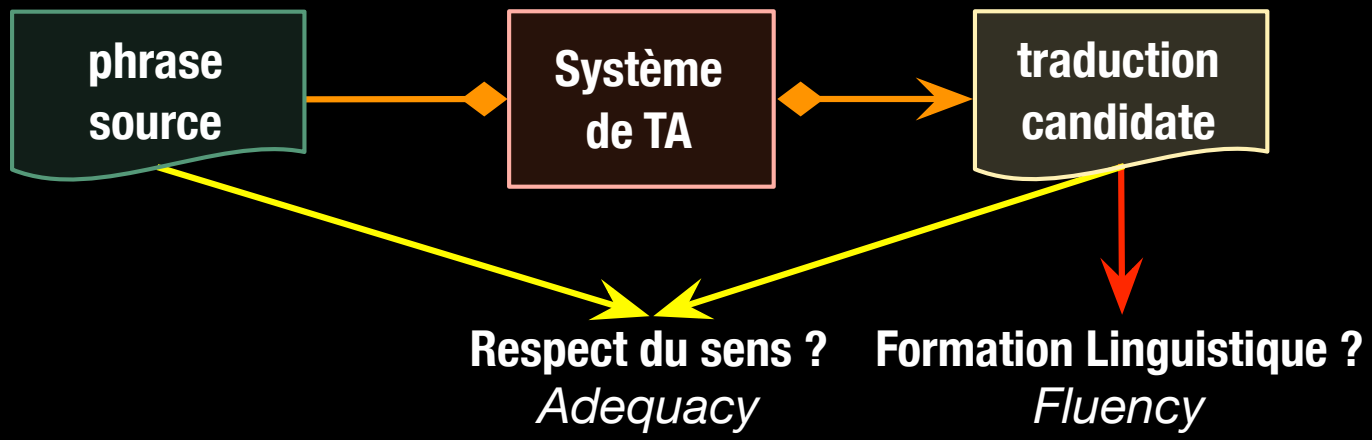
- Situation aujourd'hui
  - Ère des systèmes "minutes" (traduction statistique)
    - corpus bilingues alignés
    - boîtes à outils sur le Web + tuning
  - Ère de l'évaluation compétitive
    - NIST organise les campagnes des projets DARPA
      - les bailleurs de fonds veulent connaître les progrès
    - C-STAR organise les campagnes IWSLT

Plan

- Méthodes d'évaluation en usage
  - Subjectives (humaines)
  - Objectives (automatiques)
- Mise en œuvre et critique
- Propositions de nouvelles pratiques
  - HTER (post-édition multi-étages)
  - Distance d'édition "simple"

Évaluation subjective

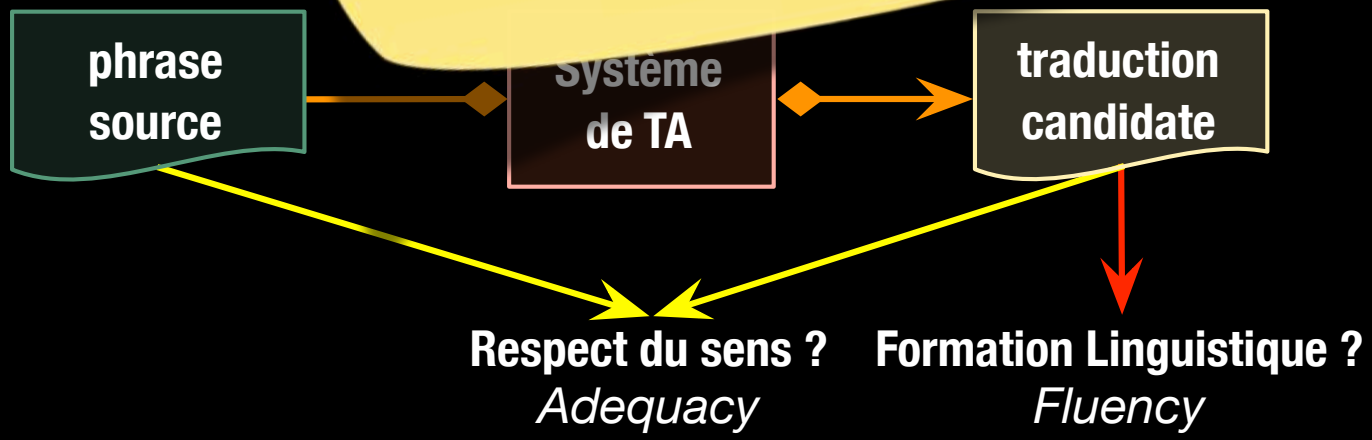
- Ancienne pratique
  - Juger la bonne formation linguistique de la traduction
  - Juger la justesse, la fidélité de la traduction ...
    - ... par rapport à la phrase source



Évaluation subjective

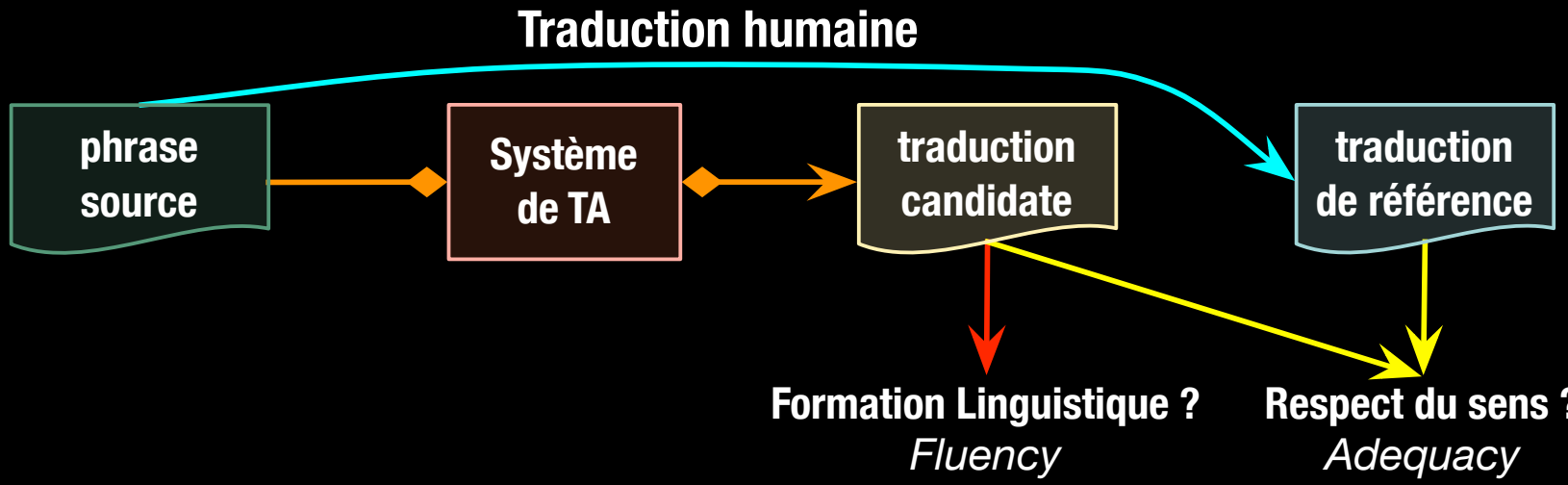
- Ancienne
  - Juger la b… la traduction
  - Juger la ju… on …
    - … par rapp…

Il faut des juges bilingues



Évaluation subjective

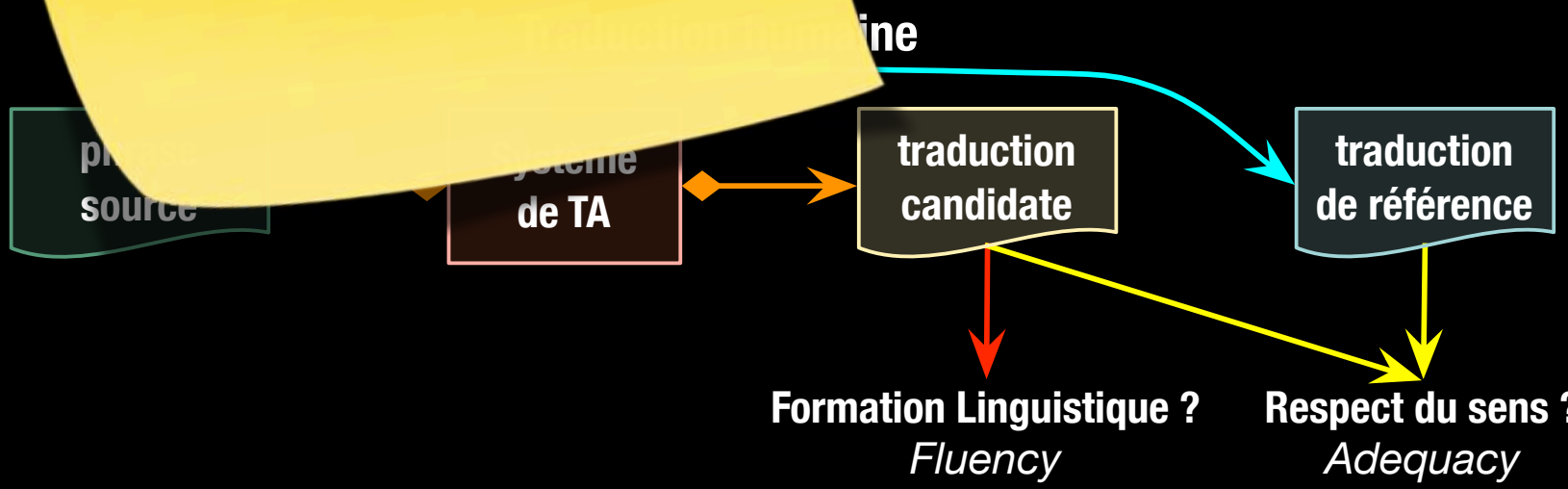
- Nouvelle pratique (NIST)
  - juger la bonne formation linguistique de la traduction
  - juger la justesse, la fidélité de la traduction ...
    - ... par rapport à une traduction de référence



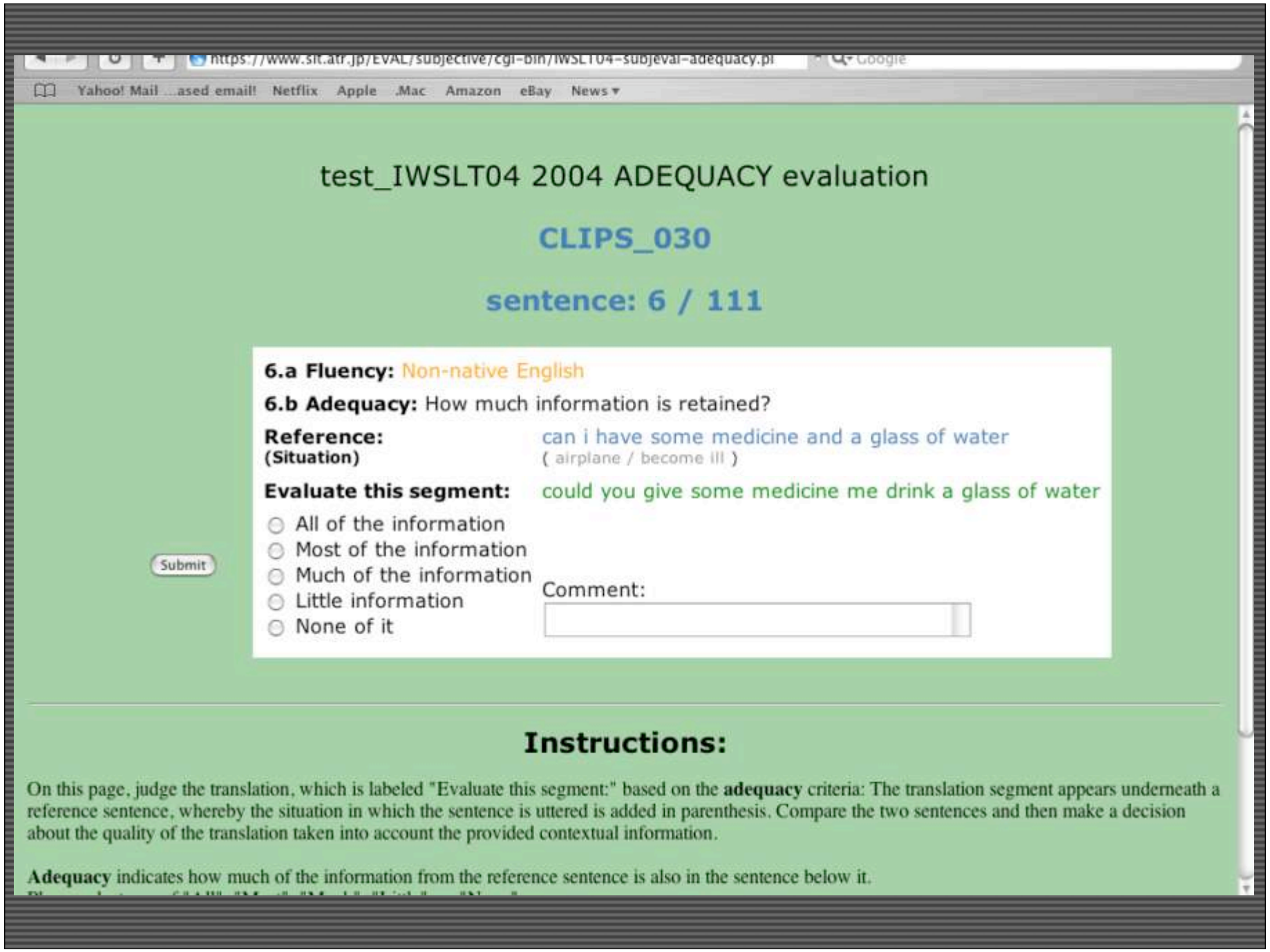
Évaluation subjective

- linguistique de la traduction
- de la traduction ...
- uction de référence

La langue source a disparu !!!!




FLUENCY evaluation form of CLIPS (file: CLIPS_030, sent: 6)
https://www.slt.atr.jp/EVAL/subjective/cgi-bin/IWSLT04-subjeval-fluency.pl
Google
Yahoo! Mail ...ased email! Netflix Apple Mac Amazon eBay News
test_IWSLT04 2004 FLUENCY evaluation
CLIPS_030
sentence: 6 / 111
6.a Fluency: How good is the English?
Evaluate this segment: could you give some medicine me drink a glass of water
Flawless English
Good English
Non-native English
Disfluent English
Incomprehensible
Comment:
Submit
Instructions:
On this page, judge the quality of the translation, which is labeled "Evaluate this segment:" based on the fluency criteria:
Fluency indicates how the evaluation segment sounds to a native speaker of English. Please select the phrase that best describes the level of English used in the translation: "Flawless English", "Good English", "Non-native English", "Disfluent English", or "Incomprehensible".
An area for comments is also provided. You may choose to leave this field blank.
A rule of thumb for grading is to spend no more than 15 seconds on each sentence. When you are done, click the "Submit" button to the left of the evaluation in order to continue.

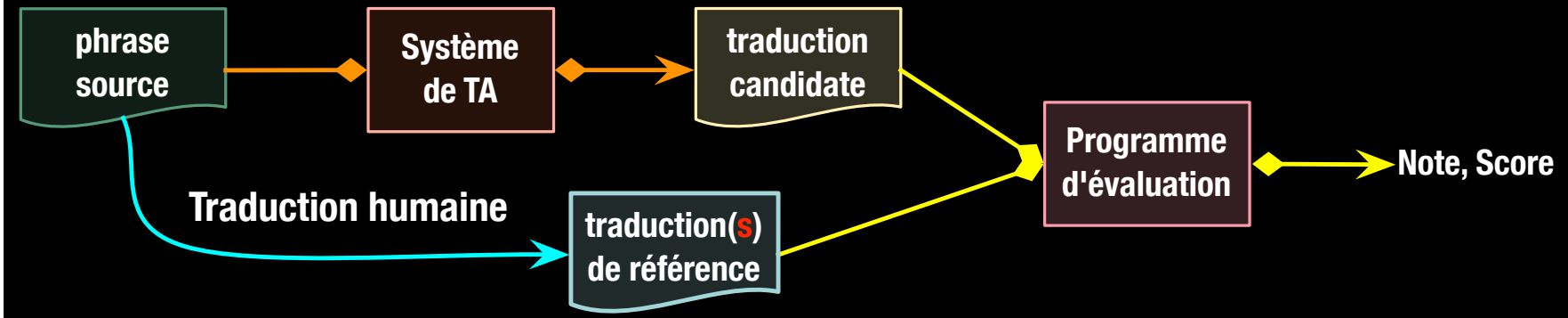

https://www.slt.atr.jp/EVAL/subjective/cgi-bin/IWSLT04-subjeval-adequacy.pl
Yahoo! Mail ...ased email! Netflix Apple .Mac Amazon eBay News
test_IWSLT04 2004 ADEQUACY evaluation
CLIPS_030
sentence: 6 / 111
6.a Fluency: Non-native English
6.b Adequacy: How much information is retained?
Reference:
(Situation)
can i have some medicine and a glass of water
(airplane / become ill)
Evaluate this segment:
could you give some medicine me drink a glass of water
All of the information
Most of the information
Much of the information
Little information
None of it
Submit
Comment:
Instructions:
On this page, judge the translation, which is labeled "Evaluate this segment:" based on the adequacy criteria: The translation segment appears underneath a reference sentence, whereby the situation in which the sentence is uttered is added in parenthesis. Compare the two sentences and then make a decision about the quality of the translation taken into account the provided contextual information.
Adequacy indicates how much of the information from the reference sentence is also in the sentence below it.

Problèmes cités par la communauté

- L'évaluation subjective demande des ressources !
  - des juges bilingues
  - des juges monolingues + des traductions de références
- L'évaluation objective prend beaucoup de temps !
- Les juges ne sont pas toujours …
  - … d'accord entre eux
  - … d'accord avec eux-même dans le temps

Idée

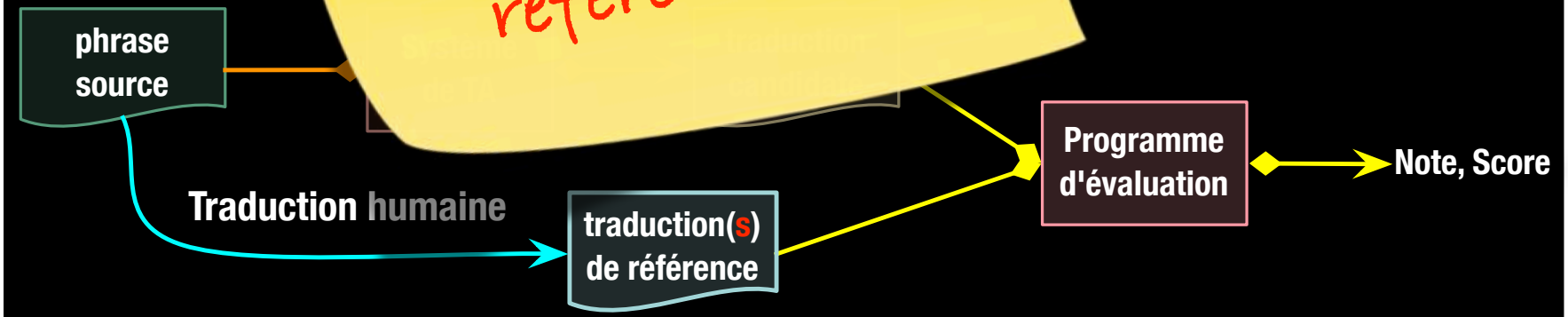
- Automatiser l'évaluation
  - Un programme déterministe qui calcule un score
    - Plus de désaccord entre juges
    - Plus de variation dans le temps
- Utiliser les données disponibles, construites pour l'évaluation subjective



Idée

- Automatiser l'évaluati
  - Un p calcule un score
    - Plu
    - Plus
- Utilise onstruites pour l'évalua

L'humain reste dans le circuit pour produire les références.



Évaluation objective

- Mesures fondées sur …
  - … la co-occurrence de n-grammes
    - BLEU [Papineni & al. 02]
    - NIST [Doddington 02]
    - METEOR [Banerjee & Lavie 05]
  - … la co-occurrence de sous-séquences (non) connexes
    - ROUGE (Famille de mesures) [lin 04]
- Mesure d'évaluation des méthodes d'évaluation
  - ORANGE [Lin 04]

BLEU

$$p_n = \frac{\sum_{C \in \{candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C \in \{candidates\}} \sum_{n\text{-}gram \in C} Count(n\text{-}gram)} \qquad \text{(B 1)}$$

BLEU

- Problèmes des traductions courtes
  - elles maximisent la précision modifiée

BLEU

- Exemple
  - Candidat : **of the**
    - Référence 1 : It is a guide to action that ensures that the military will forever heed Party command.
    - 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 direction of the Party.
  - Précision modifiée
    - monogramme : 2/2
    - bigramme : 1/1

La traduction n'est pas bonne !!!

BLEU

- Problèmes des traduction courtes
  - elles maximisent la précision modifiée
- Il faut pénaliser

$$BP_{BLEU} = \left\{ \begin{array}{ll} 1 & if \ |c| > |r| \\ e^{(1-|r|/|c|)} & if \ |c| \leq |r| \end{array} \right\} \qquad \text{(B 2)}$$

BLEU

- Calcul de BLEU

$$BLEU = BP_{BLEU} \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad \text{(B 3)}$$

$$w_n = 1/N$$

- Problèmes
  - Moyenne géométrique (w_n est une constante)
    - surévaluation des longs n-grammes moins nombreux
  - Tous les n-grammes ont le même poids

Expérience pratique

- IWSLT 2004 *(Évaluation compétitive)*
  - Système Systran
    - souvent utilisé dans les compétitions compétitives
  - Couple de langues
    - japonais ➟ anglais
  - Évaluation subjective : NIST (adéquation, fluidité)
  - Évaluation objective : BLEU, NIST, GTM, WER, PER
  - Résutat
    - 4^{ième} sur 4

Expérience pratique

- Nouvelles données
  - Traductions Systran japonais ➟ anglais révisées manuellement (de nouveaux candidats)

Expérience pratique

- Nouvelles données
  - Traductions Systran japonais ➟ anglais révisées manuellement (de nouveaux candidats)
- Nouveaux résultats
  - ⚠ Les traductions révisées se classent 3ième sur 4 !!!

Expérience pratique

- Nouvelles données
  - Traductions candidates Systran japonais ➟ anglais révisées manuellement
- Nouveaux résultats
  - ⚠ Les traductions révisées se classent 3ième sur 4 !!!

⚠ Les traductions humaines ne sont pas bonnes ?

- **Elles sont très bonnes !!!**
- …, **mais elles n'imitent pas les références !!!**

Expérience pratique

- Nouvelles données
  - Traduc… …is ➟ anglais révisées man…
- Nouv…
  - ⚠ Les … …e sur 4 !!!
- ⚠ Les tra… … bonnes ?
  - **Elles so… …!!!**
  - **…, mais elles n'imitent pas les références !!!**

cf. aussi [Callison-Burch & al. 2006] (discussion de BLEU)

Problèmes liés à ces pratiques

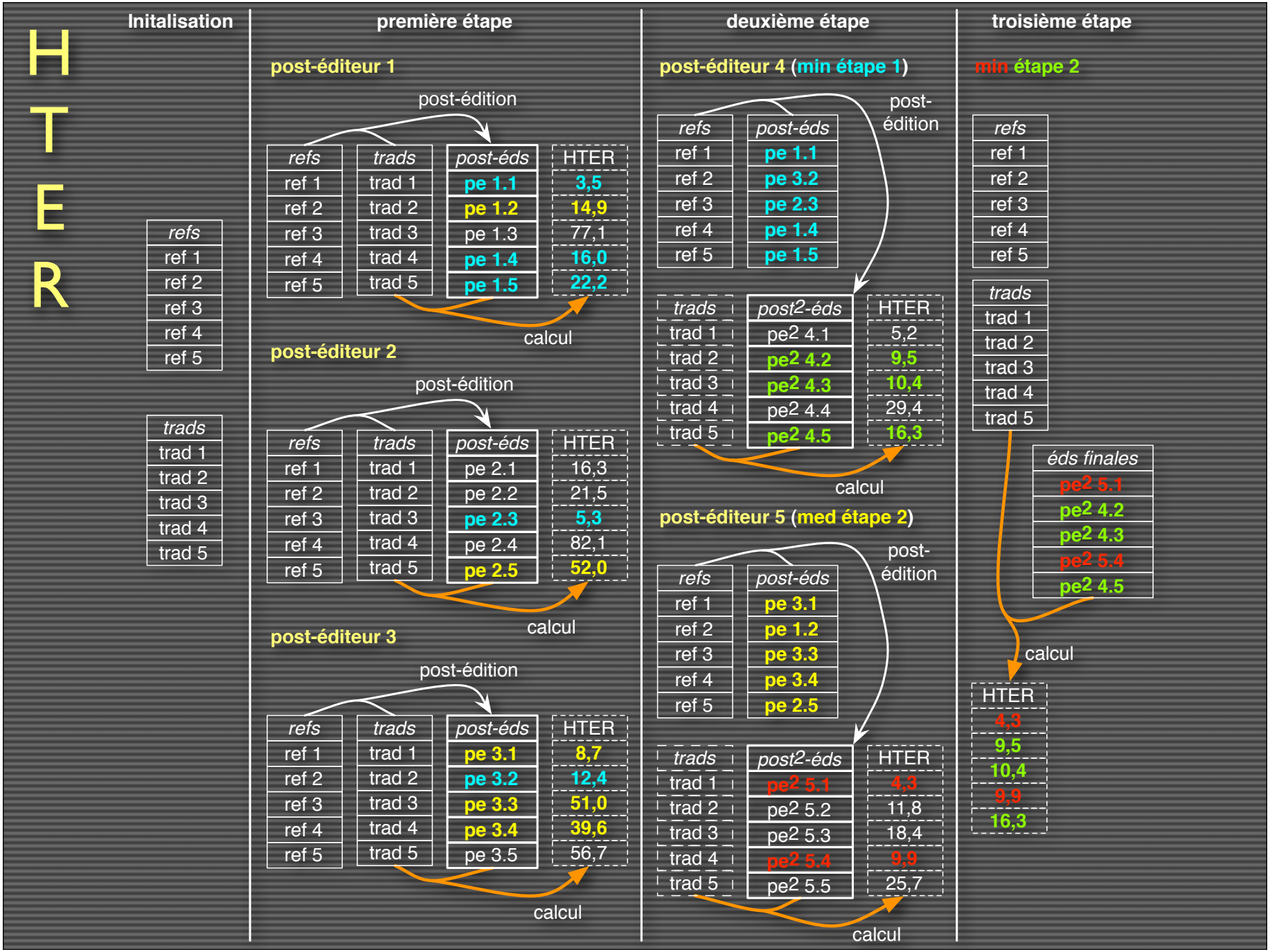
- ✓ *Absence de mesure de la qualité linguisitique*
- Faible corrélation avec les jugements subjectifs
- Mauvais scores à des traductions HQ *[Culy 03]*
- Phénomènes non gérés ? *[King]*
- Idée :
  - mesurer l'effort nécessaire à la production d'une traduction utile à la tâche (et produire de nouvelles références par effet de bord)

Problèmes liés à ces pratiques

- ✓ *Absence de mesure de l'utilisabilité pratique*
- La qualité linguistique n'est pas corrélée avec l'utilisabilité pratique
  - Systran-Euratom (1972) :
    - qualité=1/5, utilité=4.5/5
  - Les transcriptions de monologues ou de dialogue interprétés sont jugés comme de mauvaises traductions pourtant le résultat est très utile

Nouvelle pratique : HTER

- Initiée par le projet GALE (DARPA)
  - Joseph Olive ; IWSLT 05 : *"Vos mesures objectives ne veulent rien dire sur la qualité des traductions, nous ferons autrement dans GALE. Nous avons besoin de savoir si vos traductions sont utiles, ou si elles ne servent à rien."*
- Proposition HTER (Human-targeted Translation Error Rate)
  [Przybocki et al., 2006] [Snover et al., 2006]
  - Utilisation de la distance d'édition
  - Mise en œuvre en plusieurs étapes


HTER
Initalisation
refs
ref 1
ref 2
ref 3
ref 4
ref 5
trads
trad 1
trad 2
trad 3
trad 4
trad 5
première étape
post-éditeur 1
post-édition
refs
ref 1
ref 2
ref 3
ref 4
ref 5
trads
trad 1
trad 2
trad 3
trad 4
trad 5
post-éds
pe 1.1
pe 1.2
pe 1.3
pe 1.4
pe 1.5
HTER
3,5
14,9
77,1
16,0
22,2
calcul
post-éditeur 2
post-édition
refs
ref 1
ref 2
ref 3
ref 4
ref 5
trads
trad 1
trad 2
trad 3
trad 4
trad 5
post-éds
pe 2.1
pe 2.2
pe 2.3
pe 2.4
pe 2.5
HTER
16,3
21,5
5,3
82,1
52,0
calcul
post-éditeur 3
post-édition
refs
ref 1
ref 2
ref 3
ref 4
ref 5
trads
trad 1
trad 2
trad 3
trad 4
trad 5
post-éds
pe 3.1
pe 3.2
pe 3.3
pe 3.4
pe 3.5
HTER
8,7
12,4
51,0
39,6
56,7
calcul
deuxième étape
post-éditeur 4 (min étape 1)
post-édition
refs
ref 1
ref 2
ref 3
ref 4
ref 5
post-éds
pe 1.1
pe 3.2
pe 2.3
pe 1.4
pe 1.5
trads
trad 1
trad 2
trad 3
trad 4
trad 5
post2-éds
pe2 4.1
pe2 4.2
pe2 4.3
pe2 4.4
pe2 4.5
HTER
5,2
9,5
10,4
29,4
16,3
calcul
post-éditeur 5 (med étape 2)
post-édition
refs
ref 1
ref 2
ref 3
ref 4
ref 5
post-éds
pe 3.1
pe 1.2
pe 3.3
pe 3.4
pe 2.5
trads
trad 1
trad 2
trad 3
trad 4
trad 5
post2-éds
pe2 5.1
pe2 5.2
pe2 5.3
pe2 5.4
pe2 5.5
HTER
4,3
11,8
18,4
9,9
25,7
calcul
troisième étape
min étape 2
refs
ref 1
ref 2
ref 3
ref 4
ref 5
trads
trad 1
trad 2
trad 3
trad 4
trad 5
éds finales
pe2 5.1
pe2 4.2
pe2 4.3
pe2 5.4
pe2 4.5
calcul
HTER
4,3
9,5
10,4
9,9
16,3

Nouvelle pratique : HTER

- Post-édition OK
  - idée que nous reproposions déjà *[Blanchon & al. 04]*
- ⚠ Mais
  - il faut des références
  - il faut 5 post-éditeurs
  - aux étapes 2 et 3 on mélange les post-éditions des ≠ post-éditeurs ➟ risques d'incohérences
- ✓ mise en œuvre très lourde et complexe

Nouvelle pratique : distance d'édition

- Traduire avec Systran & Reverso (anglais-français) …
- … des énoncés de type livre de phrases …
- … sur 2 tâches (≈2000 tours de parole/tâche)
  - Assistance
    - demande de renseignements (itinéraire, transport, loc. voiture)
  - Restauration rapide
- Évaluation (3 juges par condition)
  - subjective *a la* NIST
  - objective avec distance d'édition *[Wagner et al., 1974]*

Nouvelle pratique : distance d'édition

- Évaluation avec SECTra-w

A pickpocket took my wallet.	*Un pickpocket a pris mon portefeuille.*	Dc=0,Dw=0 Dsent= 0	*Un pickpocket a pris mon portefeuille.*
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		
About how much would a taxi be from here?	*Au sujet de combien est-ce qu'un taxi serait d'ici ?*	Dc=17,Dw=4 Dsent= 6.6	*Combien est-ce qu'un taxi coûterait d'ici ?*
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		
About ten minutes.	*Approximativement dix minutes.*	Dc=0,Dw=0 Dsent= 0	*Approximativement dix minutes.*
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		
Actually I'm on my period.	*Réellement je suis sur ma période.*	Dc=26,Dw=6 Dsent= 10	*En fait j'ai mes règles.*
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		
Is that a problem?	*Est-ce que c'est un problème ?*	Dc=0,Dw=0 Dsent= 0	*Est-ce que c'est un problème ?*
	(A1) (A2) (A3) (A4) (A5) (F1) (F2) (F3)		

Nouvelle pratique : distance d'édition

- Évaluation objective

Phrase Dsent	Assistance	Restauration	
	Reverso	Reverso	Systran
min	0	0	0
max	15,8	25,6	21,8
moyenne	2,1	2,5	2,9
médiane	1	1,8	2,2

Nouvelle pratique : distance d'édition

- Commentaires
  - Assistance Reverso meilleur en Subjective & Objective
  - Restauration Reverso & Systran très proches
    - Ordre maintenu en objective stricte
    - Ordre inversé en objective généreuse
      - Besoin de mieux regarder les données !
  - On connaît l'effort nécessaire pour obtenir x% de traductions utiles
    - cet effort peut être réduit par des remplacements globaux
  - On dispose d'une référence pour faire de l'évaluation

plus de détails

Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche

Hervé Blanchon, Christian Boitet

TAL vol. 48-1