La préédition avec des règles peu coûteuses, utile pour la TA statistique des forums ?

Johanna Gerlach[1], Victoria Porro[1], Pierrette Bouillon[1], Sabine Lehmann[2]
(1) UNIVERSITÉ DE GENÈVE FTI/TIM - 40, bvd Du Pont-d'Arve, CH-1211 Genève 4, Suisse
(2) ACROLINX GmbH, Friedrichstr. 100, 10117 Berlin, Allemagne
Johanna.Gerlach@unige.ch, Victoria.Porro@unige.ch, Pierrette.Bouillon@unige.ch, Sabine.Lehmann@acrolinx.com

Le projet européen ACCEPT

« Automated Community Content Editing PorTal »
www.accept.unige.ch
Améliorer la traduction des forums par trois méthodes :
- préédition
- post-édition
- techniques issues de la TA statistique elle-même

Forums utilisés dans le projet :
- Symantec (forums informatiques)
- Traducteurs Sans Frontières (textes médicaux)

Technologies utilisées : Moses, Acrolinx

Questions principales

- **Est-il possible d'écrire avec la plate-forme Acrolinx des règles de préédition utiles pour la TA ?**
- **Avec quel impact ?**
- **Peut-on obtenir le même gain avec d'autres méthodes ?**
- **Quelle méthode d'évaluation pour cette tâche ?**

Exemples

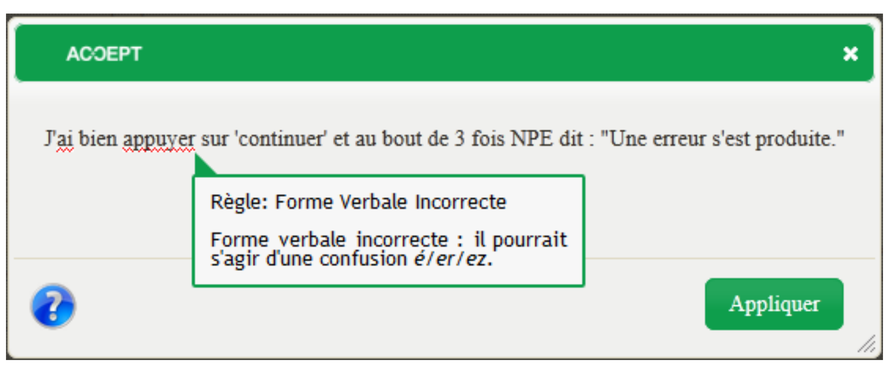
La sa ne pose pas de problème (du moins on ne recoit pas l'alerte).

La dessus on est sur de rien et c'est valable pour n'importe qu'elle antivirus que j'ai put voir, contrairement aux dires.

Règles de préédition : critères

- Se focalisent sur trois phénomènes :
  - **Confusion entre mots (liée aux homophones)**
  - **Langue informelle et familière**
  - **Différences syntaxiques entre le français et l'anglais**
- Les fautes doivent pouvoir être détectées avec le formalisme d'expression régulière d'Acrolinx.
- Les règles doivent pouvoir être appliquées facilement par les utilisateurs de forums : précision plus importante que rappel

Plug-in ACCEPT



3 ensembles de règles

Set 1 : règles (automatiques) pour humains avec suggestion unique :
- Homophones de différentes catégories syntaxiques : *a/à, la/là*
- Ponctuation (tiret, virgule, etc.) et élision

Set 2 : règles pour humains nécessitant une intervention humaine :
- Accord
- Confusion temps/mode
- Style (langage informel/familier)

Set 3 : règles pour la TA uniquement :
- Seconde personne singulier –> seconde personne pluriel
- Ordre des mots (clitiques, rien, jamais)

Exemple de règle « si et que » ⇨ « si et si »

Patron (Trigger) : @conj []{2-15} 'et' 'que'	Conjonction « si » suivie de 2 à 15 mots puis de la conjonction « et » et de « que »
Suggestion : 'que' -> @conj	« que » est remplacé par la conjonction « si »
Exception : que []+ @conj []* que	Bloquer la règle si la conjonction « si » est précédée par « que » + un ou plusieurs mots.

Impact sur la TA statistique

	Non préédité	Préédité
Set 1		
Source	oups j'ai oublié , j'ai **sa** aussi.	oups j'ai oublié, j'ai **ça** aussi.
TA	Oops I forgot, I have **its** also.	I have forgotten, I have **this** too.
Source	**avez vous** des explications ou astuces pour que cela fonctionne?	**Avez-vous** des explications ou astuces pour que cela fonctionne?
TA	**Have you** explanations or tips for it to work?	**Do you have** any explanations or tips for it to work?
Set 2		
Source	**Tu as** lu le **tuto** sur le forum?	**As-tu** lu le **tutoriel** sur le forum?
TA	**You have** read the **Tuto** on the forum?	**Have you** read the **tutorial** on the forum?
Set 3		
Source	J'ai apporté une modification dans le titre de **ton** sujet.	J'ai apporté une modification dans le titre de **votre** sujet
TA	I have made a change in the title of **tone** subject	I have made a change in the title of **your** issue
Source	Il est recommandé de **la** tester sur une machine dédiée.	Il est recommandé de tester **ça** sur une machine dédiée.
TA	It is recommended **to the test** on a dedicated machine.	It is recommended **to test it** on a dedicated machine.

Évaluation : méthode et interface

- **Évaluation humaine comparative** des traductions des phrases brutes et prééditées
- 5 jugements : first/second clearly/slightly better, about the same.
- **2 types d'évaluateurs** :
  - Travailleurs recrutés sur **Amazon Mechanical Turk** (AMT)
  - **Étudiants** en traduction de langue maternelle anglaise de la FTI

Résultats par ensemble et par type de juges

	% application	Inclus dans l'évaluation	NoImpact	**Raw Better**	About hte same	**Pre-edited Better**	No majority judgement	impact	significatif p<0.05
Ensemble 1 (précision = 91%)									
AMT	42%	611	30%	**10%**	8%	**49%**	4%	pos	oui
trad.	42%	611	30%	**9%**	14%	**42%**	5%	pos	oui
Ensemble 2 (précision = 88%)									
AMT	20%	674	15%	**17%**	6%	**58%**	4%	pos	oui
trad.	20%	674	15%	**16%**	11%	**53%**	4%	pos	oui
Ensemble 3 (précision = 98%)									
AMT	36%	448	19%	**17%**	6%	**53%**	5%	pos	oui
trad.	36%	448	19%	**15%**	12%	**50%**	4%	pos	oui

Résultats par catégorie de règles

Catégorie	Total de cas marqués	Inclus dans l'évaluation	No impact	**Raw Better**	About hte same	**Pre-edited Better**	No majority judgement	p-value	Significatif p<0.05
ponctuation	3796	416	35%	**8%**	8%	**44%**	5%	2.4E-24	oui
tu	1968	50	40%	**6%**	8%	**42%**	4%	5.2E-04	oui
clitiques	1206	150	21%	**18%**	9%	**46%**	5%	2.9E-05	oui
informel	971	367	11%	**18%**	5%	**59%**	6%	1.4E-18	oui
homophones	659	323	17%	**12%**	10%	**57%**	4%	1.4E-22	oui
grammaire (accord)	591	150	21%	**15%**	7%	**55%**	2%	7.2E-09	oui
reformulation	177	177	11%	**15%**	5%	**65%**	5%	1.3E-13	oui
ordre	71	71	17%	**24%**	4%	**48%**	7%	2.5E-02	oui
grammaire (autres)	36	28	32%	**32%**	7%	**25%**	4%	8.0E-01	non