

Classification automatique des procédés de traduction

Yuming Zhai, Gabriel Illouz, Anne Vilnat

LIMSI, CNRS, Univ. Paris-Sud, Univ. Paris-Saclay
Orsay, France

4 juillet 2019
TALN



Une bonne traduction : plein de possibilités

Traduction mot à mot

we all share the same

nous partageons tous les mêmes

Une bonne traduction : plein de possibilités

Traduction mot à mot

we all share the same
nous partageons tous les mêmes

Glissement syntaxique

the statistics you hear about
les statistiques qui nous sont communiquées

Une bonne traduction : plein de possibilités

Traduction mot à mot

we all share the same
nous partageons tous les mêmes

Glissement syntaxique

the statistics you hear about
les statistiques qui nous sont communiquées

Glissement sémantique

convert that CO₂ back into fuel to drive this process
reconvertir ce CO₂ en carburant pour alimenter ce processus

Une bonne traduction : plein de possibilités

Traduction mot à mot

we all share the same
nous partageons tous les mêmes

Glissement syntaxique

the statistics you hear about
les statistiques qui nous sont communiquées

Glissement sémantique

convert that CO₂ back into fuel to drive this process
reconvertir ce CO₂ en carburant pour alimenter ce processus

Traduction plus éloignée, glissement syntaxique et sémantique

and that scar has stayed with him for his entire life
et que, toute sa vie, il a souffert de ce traumatisme

Plan de la présentation

- 1 Contexte et travaux précédents
- 2 Corpus et annotation
- 3 Classification automatique
- 4 Résultats expérimentaux
- 5 Conclusion et perspectives

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Les catégories retenues ici

	mot-à-mot	catégorie grammaticale	sémantique	syntaxe
littéral	x			

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Les catégories retenues ici

	mot-à-mot	catégorie grammaticale	sémantique	syntaxe
littéral	x			
équivalence				

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Les catégories retenues ici

	mot-à-mot	catégorie grammaticale	sémantique	syntaxe
littéral	x			
équivalence				
transposition		x		

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Les catégories retenues ici

	mot-à-mot	catégorie grammaticale	sémantique	syntaxe
littéral	x			
équivalence				
transposition		x		
modulation			x	x

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Les catégories retenues ici

	mot-à-mot	catégorie grammaticale	sémantique	syntaxe
littéral	x			
équivalence				
transposition		x		
modulation			x	x
mod+trans		x	x	x

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Les catégories retenues ici

	mot-à-mot	catégorie grammaticale	sémantique	syntaxe
littéral	x			
équivalence				
transposition		x		
modulation			x	x
mod+trans		x	x	x
généralisation			x, ↑	

Contexte de travail

Procédés de traduction étudiés par les traducteurs humains

Traduction littérale versus d'autres procédés de traduction
(Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989)

Les catégories retenues ici

	mot-à-mot	catégorie grammaticale	sémantique	syntaxe
littéral	x			
équivalence				
transposition		x		
modulation			x	x
mod+trans		x	x	x
généralisation			x, ↑	
particularisation			x, ↓	

Contexte de travail

Manque d'exploitation explicite

Traduction automatique (SMT, NMT)

(Wu et al., 2016, Mallinson et al., 2017)

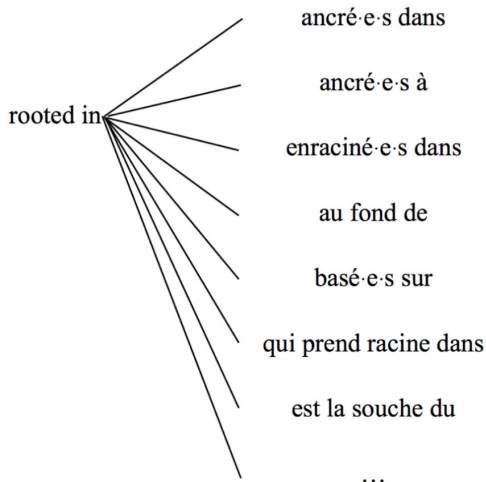
Paraphraser en exploitant des corpus parallèles bilingues

(Bannard and Callison-Burch, 2005)

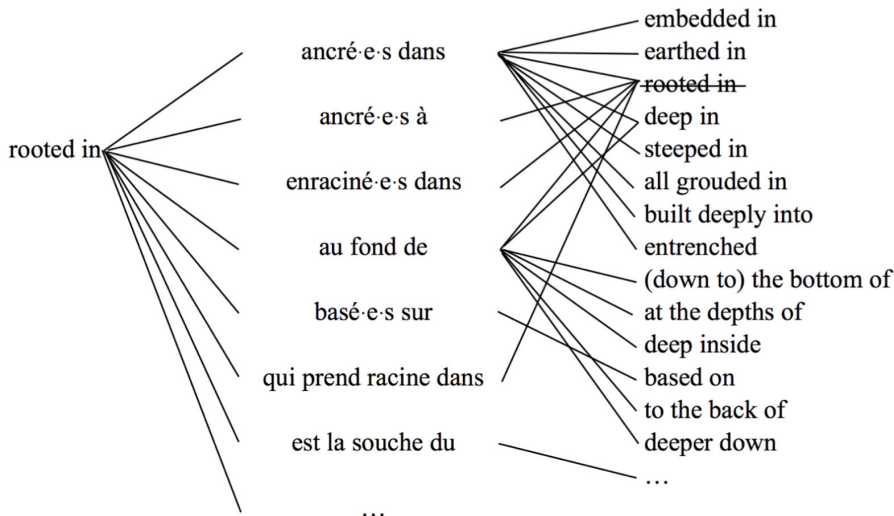
Paraphraser par équivalence de traduction (par pivot)

rooted in

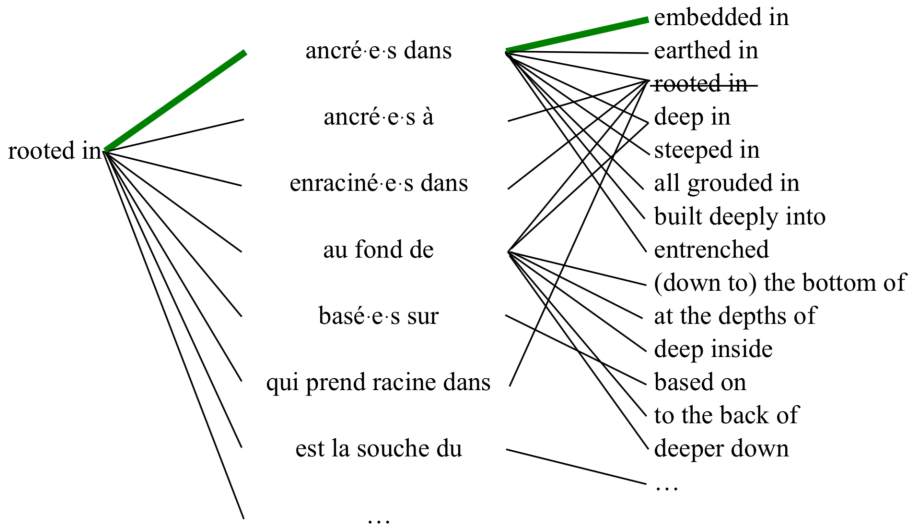
Paraphraser par équivalence de traduction (par pivot)



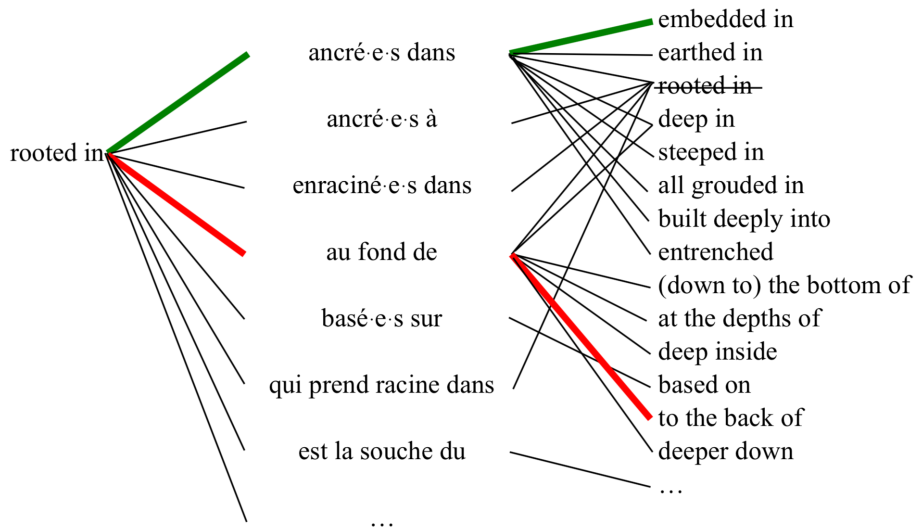
Paraphraser par équivalence de traduction (par pivot)



Paraphraser par équivalence de traduction (par pivot)



Paraphraser par équivalence de traduction (par pivot)



Relations sémantiques diverses dans PPDB 2.0

Ressource de paraphrase

PPDB (*Paraphrase Database*)

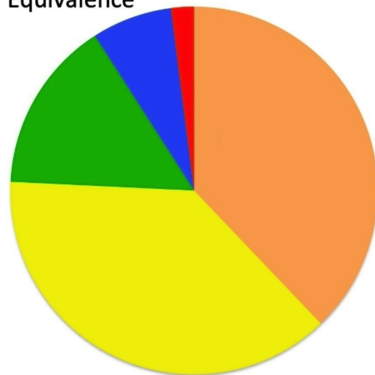
(Ganitkevitch and Callison-Burch, 2013)

Manque de contrôle sémantique

(Pavlick et al., 2015)

Équivalence : *distant* / *remote*

Équivalence



Relations sémantiques diverses dans PPDB 2.0

Ressource de paraphrase

PPDB (*Paraphrase Database*)

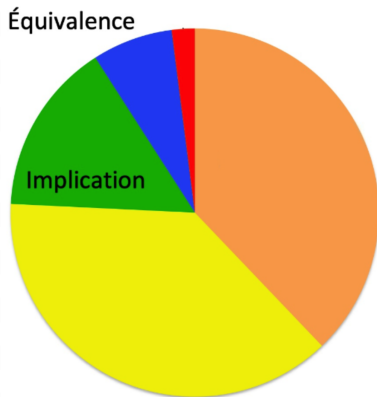
(Ganitkevitch and Callison-Burch, 2013)

Manque de contrôle sémantique

(Pavlick et al., 2015)

Équivalence : *distant* / *remote*

Implication : *tower* / *building*



Relations sémantiques diverses dans PPDB 2.0

Ressource de paraphrase

PPDB (*Paraphrase Database*)

(Ganitkevitch and Callison-Burch, 2013)

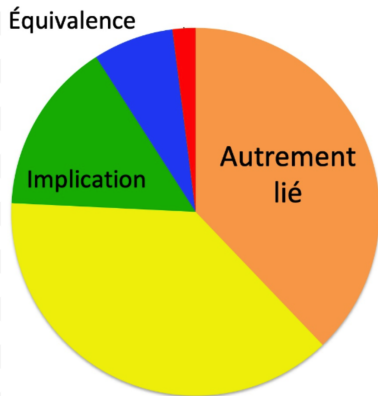
Manque de contrôle sémantique

(Pavlick et al., 2015)

Équivalence : *distant* / *remote*

Implication : *tower* / *building*

Autrement lié : *husband* / *marry to*



Relations sémantiques diverses dans PPDB 2.0

Ressource de paraphrase

PPDB (*Paraphrase Database*)
(Ganitkevitch and Callison-Burch, 2013)

Manque de contrôle sémantique

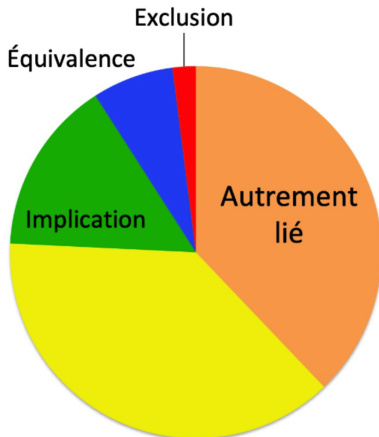
(Pavlick et al., 2015)

Équivalence : *distant* / *remote*

Implication : *tower* / *building*

Autrement lié : *husband* / *marry to*

Exclusion : *close* / *open*



Relations sémantiques diverses dans PPDB 2.0

Ressource de paraphrase

PPDB (*Paraphrase Database*)
(Ganitkevitch and Callison-Burch, 2013)

Manque de contrôle sémantique

(Pavlick et al., 2015)

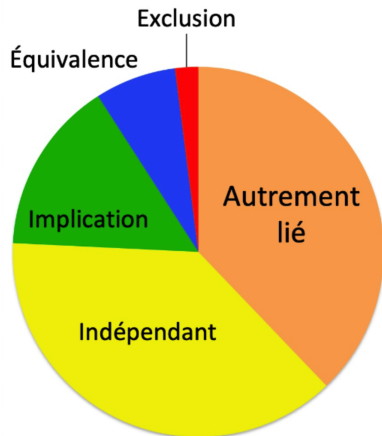
Équivalence : *distant* / *remote*

Implication : *tower* / *building*

Autrement lié : *husband* / *marry to*

Exclusion : *close* / *open*

Indépendant : *found* / *party*



Travaux précédents

Étude sur la traduction non littérale

améliorer l'alignement de mot automatique (Dorr et al., 2002)

développer une mesure de traduction littérale (Carl and Schaeffer, 2017)

comparer la traduction automatique et humaine (Lars Ahrenberg, 2017)

Travaux précédents

Étude sur la traduction non littérale

améliorer l'alignement de mot automatique (Dorr et al., 2002)

développer une mesure de traduction littérale (Carl and Schaeffer, 2017)

comparer la traduction automatique et humaine (Lars Ahrenberg, 2017)

Détection automatique des divergences en traduction

filtrer automatiquement des couples de phrases divergentes dans des corpus parallèles (Carpuat et al., 2017, Vyas et al., 2018, Pham et al., 2018)

Travaux précédents

Étude sur la traduction non littérale

améliorer l'alignement de mot automatique (Dorr et al., 2002)
développer une mesure de traduction littérale (Carl and Schaeffer, 2017)
comparer la traduction automatique et humaine (Lars Ahrenberg, 2017)

Détection automatique des divergences en traduction

filtrer automatiquement des couples de phrases divergentes dans des corpus parallèles (Carpuat et al., 2017, Vyas et al., 2018, Pham et al., 2018)

Notre contribution

- exploiter les procédés de traduction au niveau **sous-phrastique**
- classification automatique basée sur **l'annotation manuelle**
- but : meilleur contrôle sémantique dans l'extraction des paraphrases

Plan de la présentation

- 1 Contexte et travaux précédents
- 2 Corpus et annotation**
- 3 Classification automatique
- 4 Résultats expérimentaux
- 5 Conclusion et perspectives

Exemples typiques annotés (couple anglais-français)

littéral	<i>certain kinds of → certains types de</i>
équivalence	<i>alternatives → solutions de remplacement</i>
transposition	<i>astonishingly inquisitive → dotée d'une curiosité stupéfiante</i>
modulation	<i>the statistics you hear about → les statistiques qui nous sont communiquées</i>
mod+trans	<i>Steve's columns became the basis for a book → les chroniques de Steve ont inspiré un livre</i>
généralisation	<i>as we sit here in ... → alors que nous sommes à ...</i>
particularisation	<i>they have a screen → ils sont équipés d'un écran</i>

extraits du corpus *TED Talks* annoté (Zhai et al., 2018)

Statistique des données

Nombre d'instances par catégorie

littéral	3771
équivalence	289
contient_transposition (transposition+mod_trans)	342
modulation	195
généralisation	86
particularisation	215

Accord inter-annotateur

Corpus de contrôle : 100 lignes (3055 tokens anglais, 3238 tokens français)
 kappa de Cohen : **0,672**

Plan de la présentation

- 1 Contexte et travaux précédents
- 2 Corpus et annotation
- 3 Classification automatique**
- 4 Résultats expérimentaux
- 5 Conclusion et perspectives

Objectif de classification

Entrée (frontière connue) : *deceptive* → *une illusion*

Sortie : la catégorie **contient_transposition**

- sans changement de sens
- changement de catégorie grammaticale

Traits exploités

1) Analyse morpho-syntaxique

- comptage du nombre d'occurrence des étiquettes de PoS :

	ADJ	DET	NOUN	...	INTJ
anglais	1	0	0	...	0
français	0	1	2	...	0

- similarité cosinus entre ces deux vecteurs
- vérification du patron de changement de séquence de PoS
methodologically → *de façon méthodologique*
ADV → *ADP NOUN ADJ*

Traits exploités

1) Analyse morpho-syntaxique

- comptage du nombre d'occurrence des étiquettes de PoS :

	ADJ	DET	NOUN	...	INTJ
anglais	1	0	0	...	0
français	0	1	2	...	0

- similarité cosinus entre ces deux vecteurs
- vérification du patron de changement de séquence de PoS
methodologically → *de façon méthodologique*
ADV → *ADP NOUN ADJ*

2) Surface

- nombre de tokens anglais (len_e), français (len_f)
- ratio de ces nombres (len_e/len_f , len_f/len_e)
- distance Levenshtein entre deux segments

Traits exploités

3.1) Analyse syntaxique en constituant

- trait binaire (0 ou 1) :

étiquettes PoS :

adapt (*verbe*) → *adaptation* (*nom*)

étiquettes du nœud non terminal :

are a year behind (*phrase verbale*)

→ *avons une année de retard* (*phrase verbale*)

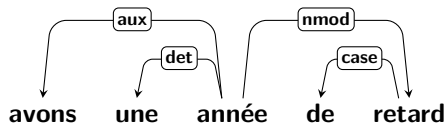
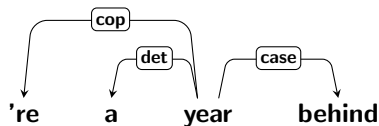
catégorie des étiquettes :

adjacent (*adjectif*) → *qui n'était pas loin* (*phrase verbale*)

Traits exploités

3.2) Analyse syntaxique en dépendance

- à l'intérieur des segments



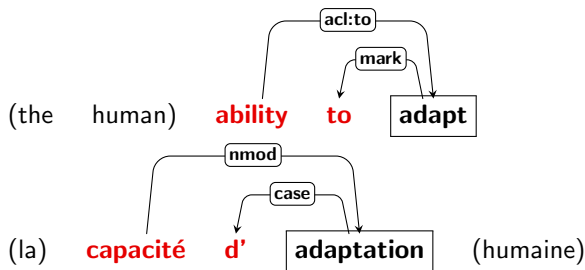
	amod	det	nmod	case	...	nsubj
anglais	0	1	0	1	...	0
français	0	1	1	1	...	0

Table: Comptage des étiquettes de relation de dépendance

Traits exploités

3.2) Analyse syntaxique en dépendance

- à l'extérieur des segments



	acl	mark	nmod	case	...	nsubj
anglais	1	1	0	0	...	0
français	0	0	1	1	...	0

Table: Comptage des étiquettes de relation de dépendance

Traits exploités

4) Ressource externe multilingue : ConceptNet (Speer et al., 2017)

- similarité cosinus entre les plongements (*ConceptNet Numberbatch*)
expressions multi-mots :

climate_change → **changement_climatique**

moyenne des plongements lexicaux sur les mots pleins :

're less burdened by → **est moins un fardeau**

Traits exploités

4) Ressource externe multilingue : ConceptNet (Speer et al., 2017)

- similarité cosinus entre les plongements (*ConceptNet Numberbatch*)
expressions multi-mots :

climate_change → **changement climatique**

moyenne des plongements lexicaux sur les mots pleins :

're less burdened by → **est moins un fardeau**

- comment la paire est liée dans la ressource

0 : lié directement

1 : lié indirectement : **complete** ← **complet** / **entier** → **totale**

2 : pas lié

Traits exploités

4) Ressource externe multilingue : ConceptNet (Speer et al., 2017)

- similarité cosinus entre les plongements (*ConceptNet Numberbatch*)
expressions multi-mots :
climate_change → **changement climatique**
moyenne des plongements lexicaux sur les mots pleins :
're less burdened by → **est moins un fardeau**
- comment la paire est liée dans la ressource
0 : lié directement
1 : lié indirectement : **complete** ← **complet / entier** → **totale**
2 : pas lié
- pourcentage des tokens liés indirectement
deceptive ← **illusoire** → une **illusion**

Traits exploités

5) Alignement de mot automatique

Table de traduction lexicale générée par *Berkeley Word Aligner*

- moyenne des entropies de traduction lexicale sur des mots pleins

Traits exploités

5) Alignement de mot automatique

Table de traduction lexicale générée par *Berkeley Word Aligner*

- moyenne des entropies de traduction lexicale sur des mots pleins
- pondération lexicale bi-directionnelle sur les mots pleins
(Koehn et al., 2003)

$$lex(e|f, A) = \prod_{i=1}^{length(e)} \frac{1}{|\{j|(i,j) \in A\}|} \sum_{\forall (i,j) \in A} w(e_i|f_j) \quad (1)$$

Traits exploités

5) Alignement de mot automatique

Table de traduction lexicale générée par *Berkeley Word Aligner*

- moyenne des entropies de traduction lexicale sur des mots pleins
- pondération lexicale bi-directionnelle sur les mots pleins
(Koehn et al., 2003)

$$lex(e|f, A) = \prod_{i=1}^{length(e)} \frac{1}{|\{j|(i,j) \in A\}|} \sum_{\forall (i,j) \in A} w(e_i|f_j) \quad (1)$$

- somme de différence de probabilités de traduction lexicale
alternatives → **alternatives** $P = 0,4$
alternatives → solutions de remplacement $P = 0,07$

Plan de la présentation

- 1 Contexte et travaux précédents
- 2 Corpus et annotation
- 3 Classification automatique
- 4 Résultats expérimentaux**
- 5 Conclusion et perspectives

Plan expérimental

- littéral (3771) vs non_littéral (5 classes, 1127)
→ différentes configurations d'évaluation
- entraîner différents classifieurs avec *Scikit-learn*
(**RandomForest**, *Multi-Layer Perceptron*, *Logistic Regression*, *Support Vector Machine*, *K-nearest Neighbors*, *Decision Tree*, *Bernoulli Naive Bayes*, *multinomial Naive Bayes*, *Gaussian Naive Bayes*)
- les hyperparamètres sont optimisés
- validation croisée à 5 plis
- baseline Dummy (prédire toujours la classe majoritaire)

Résultats expérimentaux (utilisant tous les traits)

Distribution de classes	Classifieur	Exactitude moyenne	Micro-F1	Macro-F1
Six classes				
six classes, avec 3771 <i>littéral</i>	Dummy RandomForest	76,99% 83,10% \pm 0,35%	0,77 0,83	0,14 0,44
six classes, avec 200 <i>littéral</i>	Dummy RandomForest	25,77% 57,04% \pm 1,47%	0,26 0,57	0,07 0,52

Résultats expérimentaux (utilisant tous les traits)

Distribution de classes	Classifieur	Exactitude moyenne	Micro-F1	Macro-F1
Six classes				
six classes, avec 3771 <i>littéral</i>	Dummy RandomForest	76,99% 83,10% \pm 0,35%	0,77 0,83	0,14 0,44
six classes, avec 200 <i>littéral</i>	Dummy RandomForest	25,77% 57,04% \pm 1,47%	0,26 0,57	0,07 0,52
Deux classes				
<i>littéral</i> (3) : <i>non_littéral</i> (1)	Dummy RandomForest	76,99% 90,16% \pm 0,98%	0,77 0,90	0,43 0,86
<i>littéral</i> (2) : <i>non_littéral</i> (1)	Dummy RandomForest	66,67% 88,85% \pm 0,71%	0,67 0,89	0,40 0,88
<i>littéral</i> (1) : <i>non_littéral</i> (1)	Dummy RandomForest	50,00% 87,09% \pm 2,50%	0,50 0,87	0,33 0,87

Résultats expérimentaux (utilisant tous les traits)

Distribution de classes	Classifieur	Exactitude moyenne	Micro-F1	Macro-F1
Six classes				
six classes, avec 3771 <i>littéral</i>	Dummy RandomForest	76,99% 83,10% \pm 0,35%	0,77 0,83	0,14 0,44
six classes, avec 200 <i>littéral</i>	Dummy RandomForest	25,77% 57,04% \pm 1,47%	0,26 0,57	0,07 0,52
Deux classes				
<i>littéral</i> (3) : <i>non_littéral</i> (1)	Dummy RandomForest	76,99% 90,16% \pm 0,98%	0,77 0,90	0,43 0,86
<i>littéral</i> (2) : <i>non_littéral</i> (1)	Dummy RandomForest	66,67% 88,85% \pm 0,71%	0,67 0,89	0,40 0,88
<i>littéral</i> (1) : <i>non_littéral</i> (1)	Dummy RandomForest	50,00% 87,09% \pm 2,50%	0,50 0,87	0,33 0,87
Cinq classes				
cinq classes non_littéral	Dummy RandomForest	30,35% 55,10% \pm 1,45%	0,30 0,55	0,09 0,47

Analyse

Confusion entre :

- *littéral* et *équivalence* *search history* → *historique de recherche*
- *équivalence* et *contient_transposition*
all the people in the world → *la population mondiale*
- *modulation* et les autres classes

équivalence	généralisation	particularisation	modulation	contient-transposition
0,51 ± 0,02	0,25 ± 0,09	0,56 ± 0,05	0,36 ± 0,08	0,68 ± 0,02

Table: F-mesures moyennes sur les cinq plis pour chaque procédé non littéral

Analyse

Confusion entre :

- *littéral* et *équivalence* *search history* → *historique de recherche*
- *équivalence* et *contient_transposition*
all the people in the world → *la population mondiale*
- *modulation* et les autres classes

équivalence	généralisation	particularisation	modulation	contient-transposition
0,51 ± 0,02	0,25 ± 0,09	0,56 ± 0,05	0,36 ± 0,08	0,68 ± 0,02

Table: F-mesures moyennes sur les cinq plis pour chaque procédé non littéral

Quels traits contribuent le plus ?

- classification binaire : [alignement_de_mot](#)
- classification multi-class : [analyse_PoS](#) et [analyse_syntaxique](#)

Plan de la présentation

- 1 Contexte et travaux précédents
- 2 Corpus et annotation
- 3 Classification automatique
- 4 Résultats expérimentaux
- 5 Conclusion et perspectives**

Conclusion et perspectives

Bilan :

- nouvelle tâche : classification automatique des procédés de traduction
- meilleurs résultats : RandomForest avec l'ingénierie des traits
- ouvre des pistes de recherche

Conclusion et perspectives

Bilan :

- nouvelle tâche : classification automatique des procédés de traduction
- meilleurs résultats : RandomForest avec l'ingénierie des traits
- ouvre des pistes de recherche

Amélioration :

- ▶ augmentation de la taille de données
- ▶ apport éventuel d'autres traits
(e.g. probabilité de traduction de phrase)
- ▶ analyse d'erreur pour améliorer le guide d'annotation
- ▶ extension de l'étude au couple anglais-chinois

Conclusion et perspectives

Bilan :

- nouvelle tâche : classification automatique des procédés de traduction
- meilleurs résultats : RandomForest avec l'ingénierie des traits
- ouvre des pistes de recherche

Amélioration :

- ▶ augmentation de la taille de données
- ▶ apport éventuel d'autres traits
(e.g. probabilité de traduction de phrase)
- ▶ analyse d'erreur pour améliorer le guide d'annotation
- ▶ extension de l'étude au couple anglais-chinois

Perspectives :

- adaptation au corpus parallèle sans annotation
- extraction de meilleures paraphrases

Merci pour votre attention ! Des questions ?

Contact : yuming.zhai@limsi.fr

Code : <https://github.com/YumingZHAI/ctp>

Résultats des classifieurs neuronaux

Plus de détails dans cet article : (Zhai et al., 2019, Conférence CICLing) :

Architecture	Exactitude	F1 (Littéral)	F1 (Non littéral)
Plongement de caractère initialisé aléatoirement			
CNN	59,99%	0,60	0,60
MLP	71,16%	0,71	0,71
Plongement de mot pré-entraîné (FastText)			
MLP	71,25%	0,71	0,71

Table: Classification binaire (distribution équilibrée)

Architecture	Exactitude	Micro-F1	Macro-F1
Plongement de caractère initialisé aléatoirement			
CNN	34,08%	0,34	0,20
MLP	40,74%	0,41	0,34
Plongement de mot pré-entraîné (FastText)			
MLP	43,22%	0,43	0,34

Table: Classification multi-class (cinq classes non littéral)

Contribution des traits (classification binaire 1:1)

Trait	F1 Moyenne (Littéral)	F1 Moyenne (Non littéral)
pos_vecteur_comptage	0,78	0,75
posCosinus_tous_les_mots	0,69	0,69
posCosinus_mots_pleins	0,68	0,64
pos_changement_patron	0,70	0,34
distance_Levenshtein	0,75	0,75
(ratio)_longueur_token	0,74	0,71
analyse_constituant	0,66	0,50
analyse_dépendance_interne	0,76	0,73
analyse_dépendance_externes	0,61	0,62
ConceptNet_Embedding	0,73	0,73
ConceptNet_lien	0,70	0,78
ConceptNet_pourcentage_indirect	0,70	0,62
différence_probabilité_traduction	0,77	0,78
entropie_traduction	0,60	0,62
pondération_lexicale	0,78	0,80
analyse_PoS	0,78	0,75
surface	0,72	0,70
analyse_syntactique	0,76	0,76
ressource_ConceptNet	0,77	0,78
alignement_de_mot	0,84	0,85
tous les traits	0,87	0,87
tous les traits - les traits en vert	0,87	0,88

Contribution des traits (5 classes non littéral)

Trait	F1(micro-av)	F1(macro-av)	F1 (E)	F1 (G)	F1 (P)	F1 (M)	F1 (T)
pos_vecteur_comptage	0,52	0,44	0,51	0,16	0,54	0,34	0,65
posCosinus_tous_les_mots	0,39	0,29	0,25	0,00	0,45	0,23	0,54
posCosinus_mots_pleins	0,39	0,25	0,06	0,00	0,42	0,19	0,58
pos_changement_patron	0,37	0,20	0,43	0,00	0,00	0,00	0,56
distance_Levenshtein	0,36	0,26	0,32	0,00	0,31	0,18	0,49
(ratio)_longueur_token	0,39	0,34	0,36	0,21	0,39	0,26	0,49
analyse_constituant	0,36	0,19	0,00	0,00	0,41	0,00	0,56
analyse_dépendance_interne	0,45	0,39	0,37	0,21	0,47	0,31	0,59
analyse_dépendance_externe	0,35	0,26	0,27	0,05	0,33	0,15	0,51
ConceptNet_Embedding	0,32	0,27	0,28	0,02	0,38	0,25	0,40
ConceptNet_lien	0,32	0,18	0,13	0,00	0,30	0,00	0,46
ConceptNet_pourcentage_indirect	0,29	0,16	0,22	0,00	0,00	0,18	0,42
différence_probabilité_traduction	0,38	0,32	0,35	0,13	0,38	0,26	0,50
entropie_traduction	0,35	0,27	0,36	0,00	0,39	0,15	0,44
pondération_lexicale	0,32	0,24	0,36	0,02	0,28	0,15	0,41
analyse_PoS	0,51	0,43	0,48	0,15	0,52	0,34	0,64
surface	0,38	0,34	0,36	0,23	0,37	0,27	0,47
analyse_syntaxique	0,48	0,40	0,38	0,18	0,52	0,30	0,63
ressource_ConceptNet	0,34	0,28	0,24	0,02	0,46	0,23	0,44
alignement_de_mot	0,45	0,38	0,44	0,15	0,51	0,26	0,54
pos + surface + syntaxique	0,54	0,47	0,49	0,28	0,54	0,38	0,67
tous - ConceptNet	0,54	0,47	0,51	0,27	0,54	0,37	0,67
tous les traits	0,55	0,47	0,50	0,25	0,55	0,37	0,67

Résultats des classifieurs entraînés

Algorithme	Exactitude Moyenne	F1 (Littéral)	F1 (Non littéral)
RandomForest	87,09%	0,87	0,87
MLP	85,01%	0,85	0,85
LogisticRegression	84,78%	0,85	0,85
rbfSVC	85,14%	0,84	0,86
MultinomialNB	80,83%	0,81	0,81
KNN	83,41%	0,83	0,83
BernoulliNB	81,14%	0,80	0,82
DecisionTree	79,95%	0,79	0,81
GaussianNB	64,51%	0,73	0,50

Table: Classification binaire (distribution équilibrée), utilisant tous les traits

Pondération lexicale sur les mots pleins (Koehn et al., 2003)

Exemple :

astonishingly inquisitive \rightarrow **dotée d' une curiosité stupéfiante**

$$\text{lex}(e|f, A) = \prod_{i=1}^{\text{length}(e)} \frac{1}{|\{j|(i,j) \in A\}|} \sum_{\forall (i,j) \in A} w(e_i|f_j)$$

e : segment source

f : segment cible

A : l'ensemble des mots alignés

$w(e_i|f_j)$: probabilité de traduction lexicale

Plus de détails sur la ressource PPDB (Pavlick et al., 2015)

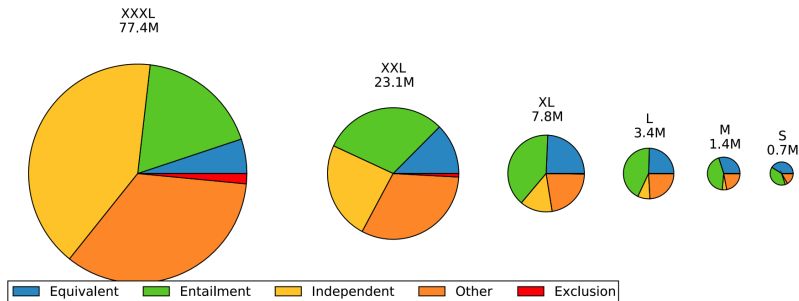


Figure 2: Distribution of entailment relations in different sizes of PPDB. Distributions are estimated from our manual annotations of randomly sampled pairs. PPDB-XXXL contains over 77MM paraphrase pairs (where the majority type is independent), compared to only 700K in PPDB-S (where the majority type is equivalent).

Annotation via l'outil Yawat (Germann, 2008)

for some time I have been interested in the placebo effect , which might seem like an odd thing for a magician to be interested in , unless you think of it in the terms that I do , which is , `` something fake is believed in enough by somebody that it becomes something real . "

pendant un temps , je me suis intéressé à l' effet placebo , ce qui peut paraître bizarre pour un magicien comme sujet de prédilection , à moins que vous ne regardiez la chose comme moi : soit " quelque chose de faux peut être tellement crédible que ça devient quelque chose de vrai " .