Construction of a Multilingual Corpus Annotated with Translation Relations

Yuming Zhai, Gabriel Illouz, Anne Vilnat

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

30 novembre 2018

Example of understanding a foreign segment

Sentence context : supply resources to human beings.

- 换作90亿人那真是捉襟见肘了。
- **Literal** translation in English: *For nine billion people, it's going to be exposing one's elbow when drawing tight the lapel of one's jacket.*
- **Literal** translation in French: *Pour neuf milliards de personnes, cela va être laisser voir son coude en tirant le revers de sa veste.*

Example of understanding a foreign segment

Sentence context : supply resources to human beings.

- 换作90亿人那真是捉襟见肘了。
- **Literal** translation in English: *For nine billion people, it's going to be exposing one's elbow when drawing tight the lapel of one's jacket.*
- **Literal** translation in French: *Pour neuf milliards de personnes, cela va être laisser voir son coude en tirant le revers de sa veste.*
- **Non-literal** translation in English: *It's going to be a stretch to do it for nine billion people.*
- **Non-literal** translation in French: *Ça sera d'autant plus difficile de le faire pour 9 milliards de personnes.*

Research Application

How to help foreign language learners to:

- understand difficult segments during their reading
- rewrite the segments that they consider as imperfect

Research Problem

Rephrase a foreign segment to help learners to understand or to rewrite, by exploiting the translation equivalence. (Bannard and Callison-Burch, 2005).

Research Problem

Rephrase a foreign segment to help learners to understand or to rewrite, by exploiting the translation equivalence. (Bannard and Callison-Burch, 2005).

Based on the algorithm of random walk and of hitting time (Kok and Brockett, 2010), our hypothesis is that **categorizing the translation relation** (*literal* versus *other translation techniques*) allows to better guide the search of reformulations in bilingual parallel corpus:

- bring a better semantic control
- bring more reformulation varieties
- make the search path interpretable

Presentation plan

1. Background
2. Ongoing work
   - Corpus construction
   - Automatic classification of translation relations
3. Prototype to be developed
4. Conclusion and perspectives

Background

Translation relations studied by human translators

Literal translation versus other translation techniques (Vinay and Darbelnet, 1958; Newmark, 1981, 1988; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002)

Background

Translation relations studied by human translators

Literal translation versus other translation techniques (Vinay and Darbelnet, 1958; Newmark, 1981, 1988; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002)

Lack of explicit exploitation

Machine translation (Statistical or Neural) (Wu *et al.*, 2016; Mallinson *et al.*, 2017)

Paraphrasing by exploiting bilingual parallel corpus (Bannard and Callison-Burch, 2005)

Background

Translation relations studied by human translators

Literal translation versus other translation techniques (Vinay and Darbelnet, 1958; Newmark, 1981, 1988; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002)

Lack of explicit exploitation

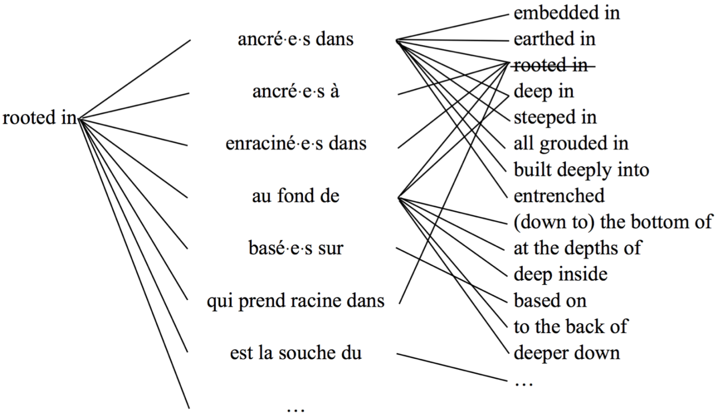
Machine translation (Statistical or Neural) (Wu *et al.*, 2016; Mallinson *et al.*, 2017)

Paraphrasing by exploiting bilingual parallel corpus (Bannard and Callison-Burch, 2005)

Research goal

- exploit the translation techniques
- have a better semantic control in the reformulations from bilingual parallel corpus

Paraphrase by exploiting translation equivalence



rooted in

ancré·e·s dans

ancré·e·s à

enraciné·e·s dans

au fond de

basé·e·s sur

qui prend racine dans

est la souche du

...

embedded in

earthed in

~~rooted in~~

deep in

steeped in

all grouded in

built deeply into

entrenched

(down to) the bottom of

at the depths of

deep inside

based on

to the back of

deeper down

...

Figure: Obtain the paraphrases for "*rooted in*" via pivot French translations.

Paraphrase by exploiting translation equivalence

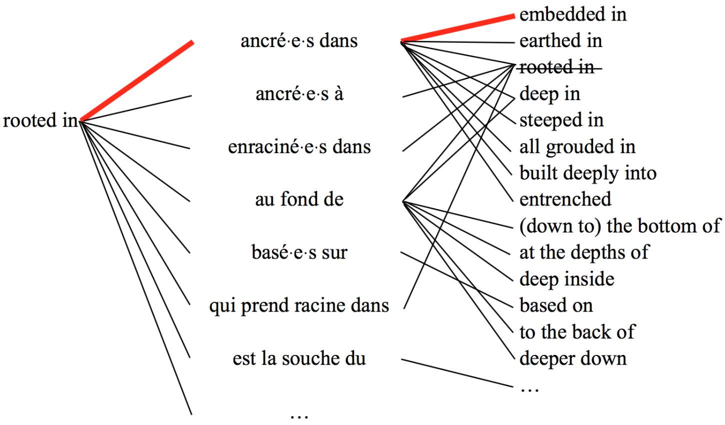


Figure: Obtain the paraphrases for "*rooted in*" via pivot French translations.

Paraphrase by exploiting translation equivalence

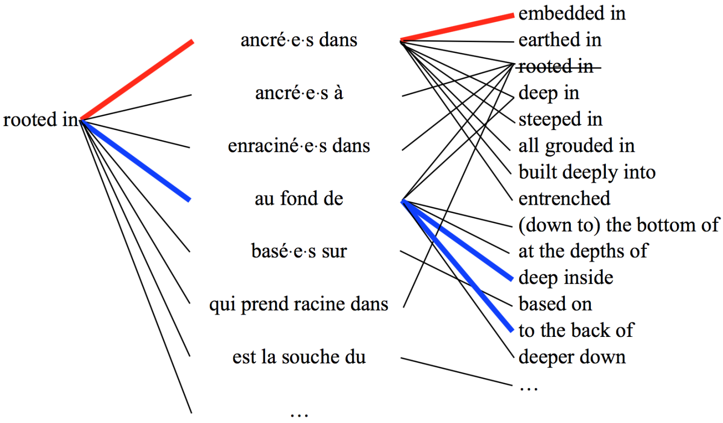


Figure: Obtain the paraphrases for "*rooted in*" via pivot French translations.

Diverse semantic relations in PPDB 2.0

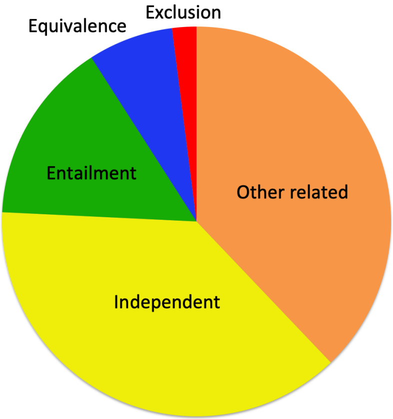


Figure: Estimated semantic relations in PPDB 2.0 XXXL (English)

Paraphrase resource

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

Lack of semantic control

Equivalence: *distant / remote*
Exclusion: *close / open*
Other related: *husband / marry to*
Independent: *found / party*
Entailment: *tower / building*
(Pavlick *et al.*, 2015)

Presentation plan

1. Background
2. Ongoing work
   - Corpus construction
   - Automatic classification of translation relations
3. Prototype to be developed
4. Conclusion and perspectives

Presentation plan

1. Background
2. Ongoing work
   - Corpus construction
   - Automatic classification of translation relations
3. Prototype to be developed
4. Conclusion and perspectives

Corpus construction

Working goal

- establish a hierarchy of translation relations to categorize them for English-French and English-Chinese
- provide a data set to train an automatic classifier of translation relations

Working method

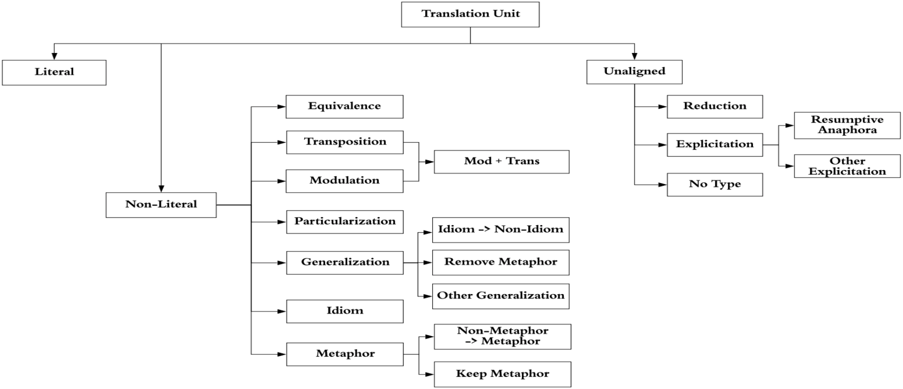
- based on the theories of translation techniques (Vinay and Darbelnet, 1958; Newmark, 1981, 1988; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002)
- annotate and analyze a multilingual parallel corpus

Corpus description

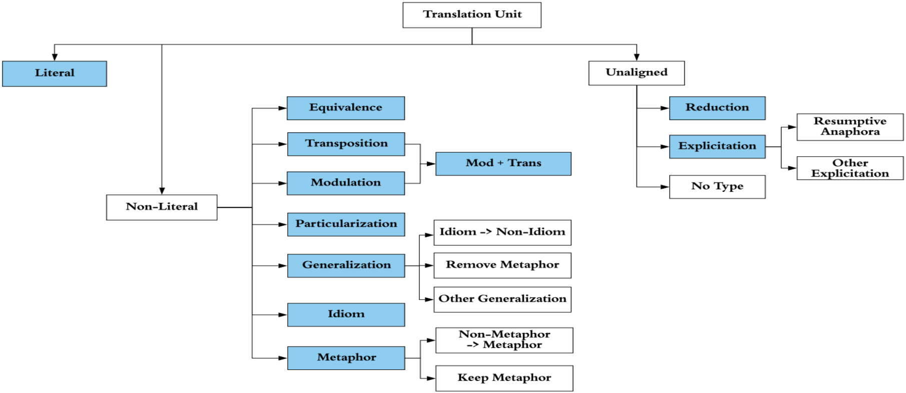
Content	transcriptions and human translations of TED Talks
Topics covered	technology, psychology, culture, science, *etc.*
# Presentations	19
Source language	English
Target language	French, Chinese
Alignment	trilingual parallel sentences
# Lines	2 436
# Lines annotated	EN-FR (925), EN-ZH (522)
# Tokens or characters (ZH)	English (51 930) French (53 932) Chinese (84 484)*

Table: Information about the annotated corpus.

Hierarchy of translation relations



Hierarchy of translation relations



(The blue labels are our annotation categories.)

Typical examples attested in our corpus (1)

Literal:
certain ***kinds*** *of* → *certains* ***types*** *de*
hatpin → ***épingle à chapeau***
Equivalence:
Birds of a feather flock together. → **Qui se ressemble s'assemble.**
if you'll pardon the pun → **si vous me passez** ce calembour
Transposition:
astonishingly inquisitive → ***dotée d'une curiosité stupéfiante***
Modulation:
that scar has stayed with him → ***il a souffert de ce traumatisme***
Mod+Trans:
this is a completely ***unsustainable*** *pattern* →
il est absolument ***impossible de continuer sur*** *cette tendance*

Typical examples attested in our corpus (2)

Particularization:

*they **have** a screen and a wireless radio → ils **sont équipés d'**un écran et d'une radio sans fil*

Generalization:

*as we **sit here** today in ... → alors que nous **sommes** à ... aujourd'hui*

Idiom:

at any given moment → à un instant "t"

Metaphor:

*if you **faint** easily → si vous **tombez dans les pommes** facilement*

Unaligned - Explicitation:

*feel their past in the wind → ressentent leur passé **souffler** dans le vent*

Unaligned - Reduction:

*look **carefully** at the area → regardez le secteur*

Typical examples attested in our corpus (3)

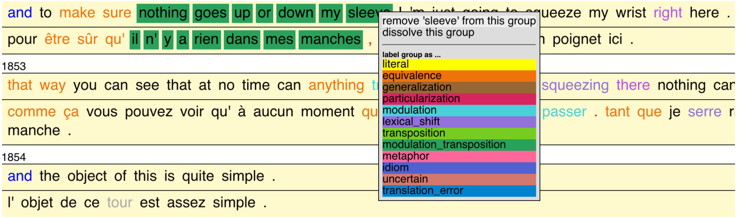
Unaligned - No type:
*minus 271 degrees, colder than → moins 271 degrés, **ce qui est** plus froid que*
Lexical shift:
*when you do a web search **for** images → quand on fait une recherche web **sur** des images*
Translation error:
*is not going to be **remembered** for its wars → ne sera pas **reconnu** pour ses guerres*
Uncertain:
Not sure about which category to assign, need more discuss.

Annotation

Annotation guidelines

- annotation conventions
- encourage using external resources (Wiktionary, Linguee, *etc.*)
- definitions, typical examples, counterexamples, borderline examples

Annotation tool: application Web *Yawat* (Germann, 2008)



and to make sure nothing goes up or down my sleeve I 'm just going to squeeze my wrist right here .

pour être sûr qu' il n' y a rien dans mes manches , … n poignet ici .

1853

that way you can see that at no time can anything t… squeezing there nothing can

comme ça vous pouvez voir qu' à aucun moment qu… passer . tant que je serre r… manche .

1854

and the object of this is quite simple .

l' objet de ce tour est assez simple .

The annotator should also decide the boundary and the word alignment.

Control study

- control corpus: 100 lines of parallel sentences. Independent annotation by two annotators for each language pair.
- inter-annotator agreement: Cohen's Kappa (Cohen, 1960)

Control study

- control corpus: 100 lines of parallel sentences. Independent annotation by two annotators for each language pair.
- inter-annotator agreement: Cohen's Kappa (Cohen, 1960)

	κ	%EN tokens	nb EN tokens
strict	0.672	72.60%	2 217
flexible	0.617	85.56%	2 613

Table: EN-FR inter-annotator agreement.

	κ	%EN tokens	nb EN tokens
strict	0.610	52.76%	1 611
flexible	0.600	74.10%	2 263

Table: EN-ZH inter-annotator agreement.

Control study

- control corpus: 100 lines of parallel sentences. Independent annotation by two annotators for each language pair.
- inter-annotator agreement: Cohen's Kappa (Cohen, 1960)

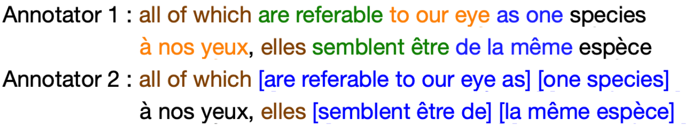
	κ	%EN tokens	nb EN tokens
strict	0.672	72.60%	2 217
flexible	0.617	85.56%	2 613

Table: EN-FR inter-annotator agreement.

	κ	%EN tokens	nb EN tokens
strict	0.610	52.76%	1 611
flexible	0.600	74.10%	2 263

Table: EN-ZH inter-annotator agreement.

- incompatible segmentation: overlap of boundaries



Annotator 1 : all of which are referable to our eye as one species
à nos yeux, elles semblent être de la même espèce

Annotator 2 : all of which [are referable to our eye as] [one species]
à nos yeux, elles [semblent être de] [la même espèce]

Annotation process in three passes

- converge both on segment boundary and on translation relation
- more time consuming but necessary for the targeted quality

	Annotator1	Annotator2
pass1	Segmentation, Alignment, Categorization	-
pass2	-	Verification, Modification
pass3	Annotator1 reviews and aims to reach consensus with annotator2.	

Table: Annotation scheme with three passes.

Annotation process in three passes

- converge both on segment boundary and on translation relation
- more time consuming but necessary for the targeted quality

	Annotator1	Annotator2
pass1	Segmentation, Alignment, Categorization	-
pass2	-	Verification, Modification
pass3	Annotator1 reviews and aims to reach consensus with annotator2.	

Table: Annotation scheme with three passes.

Enrich the annotation guide

- control study: confusion matrix of annotation; provide better definition for difficult categories
- three passes: differences between the passes show annotation difficulties

Annotation statistics (825 lines of EN-FR sentences)

Translation Relation	# raw instances	# unique instances	# EN tokens annotated (raw)	percentage token
Aligned segments				
Literal	11 771	3 718	13 228	73.80%
Equivalence	365	345	835	4.66%
Generalization	82	81	181	1.01%
Particularization	227	217	362	2.02%
Modulation	252	244	752	4.20%
Transposition	308	293	696	3.88%
Mod+Trans	53	53	188	1.05%
Idiom	2	2	6	0.03%
Metaphor	8	8	21	0.12%
Lexical_shift	398	240	475	2.65%
Translation_error	30	26	37	0.21%
Uncertain	74	74	143	0.80%
Unaligned segments				
Explicitation	175	135	-	-
Reduction	213	157	320	1.79%
No_type	-	-	680	3.79%
Total			17 924	100%

Presentation plan

1. Background

2. Ongoing work
   - Corpus construction
   - Automatic classification of translation relations

3. Prototype to be developed

4. Conclusion and perspectives

Features exploited for the automatic classification

Objective of classification

Input: *deceptive* → *une illusion* (boundary is given)
Output: category *Transposition*

Features exploited (for EN-FR):

- POS Sequence:

	ADJ	DET	NOUN	...	INTJ
English	1	0	0	0	0
French	0	1	2	0	0

- Length features: len(en), len(fr), len(en)/len(fr), len(fr)/len(en), Levenshtein distance
- Difference on the numbers of syntactic dependencies in context: outside and inside the segments
- Compare the MWE embeddings or the average of embeddings on content words (*ConceptNet Numberbatch*) (Speer et al., 2017)

Features exploited for the automatic classification

- Specific feature for detecting *Transposition*:
  - percentage of derivation links in *ConceptNet* (Speer et al., 2017):
  deceptive → ***illusoire*** ← *illusion*
  - multiword expressions linked in *ConceptNet*
  - pattern of changing the POS sequence:
  'in this simple way → *simplement de cette manière'*
  'ADP DET ADJ NOUN' → 'ADV ADP DET NOUN'
- Average word translation entropy on content words
- Bidirectional lexical weighting on content words
- For each lemmatized word, the difference of translation probability between the actual translation with the most literal translation
  actual tranlation: *being away* → *s'éloigner*
  the most literal translation: *being away* → *être loin*

Data set for binary classification

Translation Relation	Both lexical	Both MWE	Other cases	Total	Classes
Equivalence	27	219	99	345	Equivalence (**345**)
Generalization	21	20	40	81	Generalization (**81**)
Particularization	77	62	78	217	Particularization (**217**)
Modulation	57	149	38	244	Modulation (**244**)
Transposition	9	168	116	293	Contain_Transposition (**346**)
Mod+Trans	0	43	10	53	
Total					1233

Table: Statistics of non-literal translation instances. MWE: multiword expression.

Translation Relation	Both lexical	Both MWE	Other cases	Classes
Literal_random	764	193	276	Literal (1233)
Non_literal	191	661	381	Non_literal (1233)
Total				2466

Table: Statistics for binary classification.

Preliminary results (1)

Algorithm	% Mean accuracy	F1 (Literal)	F1 (Non-literal)
DummyClassifier	52.35	0.52	0.53
RandomForest	**84.71**	**0.85**	**0.85**
Multi-layer Perceptron	84.43	0.84	0.85
LogisticRegression	83.54	0.83	0.84
LinearSVM	83.25	0.83	0.84
MultinomialNB	82.32	0.80	0.84
RbfSVM	81.71	0.80	0.83
KNN	80.38	0.80	0.81
BernoulliNB	80.05	0.80	0.81
DecisionTree	78.55	0.79	0.78
GaussianNB	77.71	0.78	0.78

Table: Cross-validation with all features for binary classification.

	Literal	Non-literal
Literal	**1047**	186
Non-literal	191	**1042**

Table: Confusion matrix of RandomForest using all features. Row: gold label, column: prediction.

Presentation plan

1. Background
2. Ongoing work
   - Corpus construction
   - Automatic classification of translation relations
3. Prototype to be developed
4. Conclusion and perspectives

Prototype description

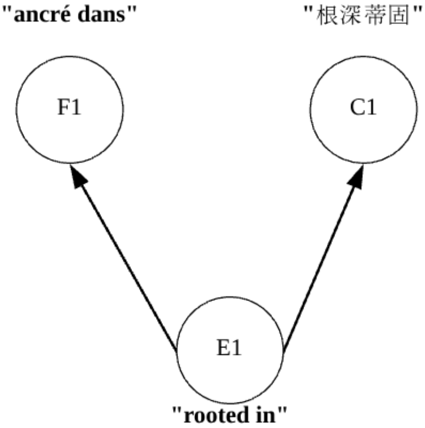


Figure: Machine translation: SMT (Gong, 2014), NMT (Mallinson *et al.*, 2017)

Prototype description

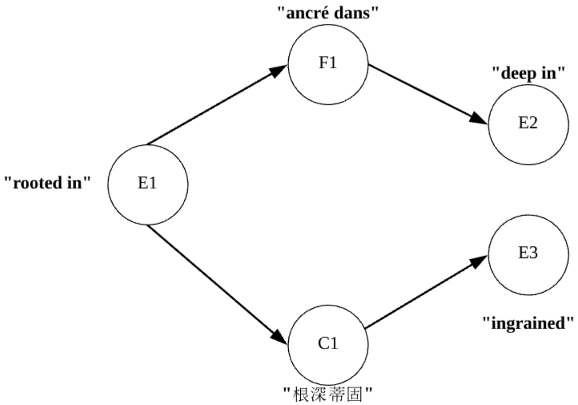


Figure: Reformulate by exploiting translation equivalence: (Bannard and Callison-Burch, 2005), (Mallinson *et al.*, 2017)

Prototype description

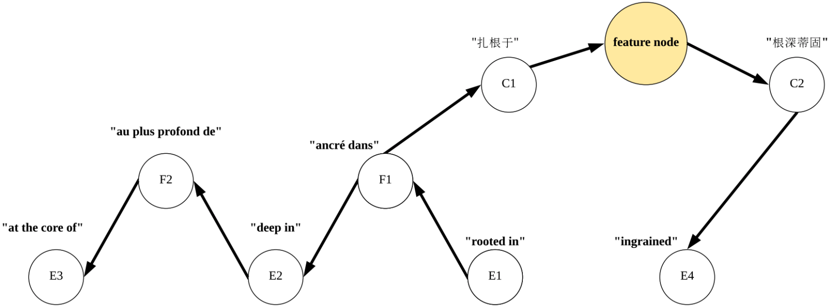


Figure: Reformulate by exploiting translation equivalence, based on the algorithm of random walk and of hitting time: (Kok and Brockett, 2010)

Prototype description

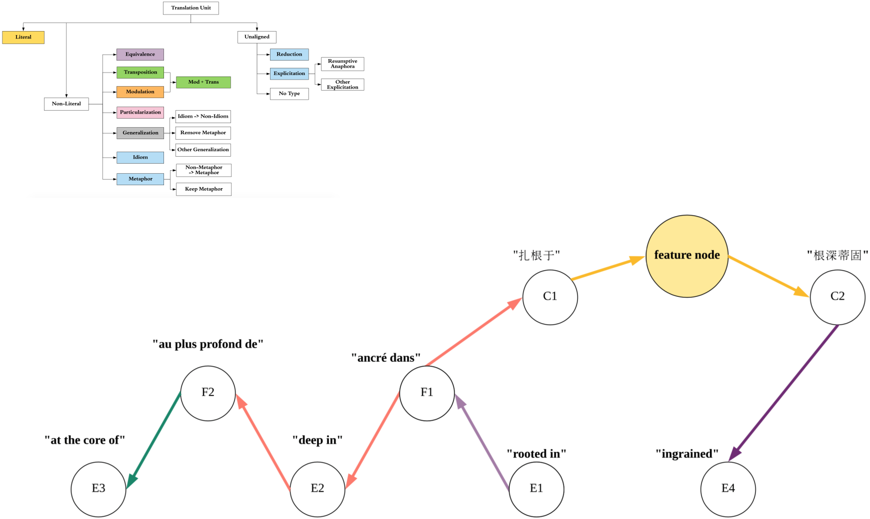


Figure: Integrate the classification of translation relations.

Presentation plan

1. Background
2. Ongoing work
   - Corpus construction
   - Automatic classification of translation relations
3. Prototype to be developed
4. Conclusion and perspectives

Conclusion and perspectives

- Propose a hierarchy of translation relations
- Annotate a parallel multilingual corpus of TED Talks (English-French, English-Chinese)
- Adopt an annotation scheme with three passes to guarantee a better annotation quality
- Preliminary results on binary classification

Conclusion and perspectives

- Propose a hierarchy of translation relations
- Annotate a parallel multilingual corpus of TED Talks (English-French, English-Chinese)
- Adopt an annotation scheme with three passes to guarantee a better annotation quality
- Preliminary results on binary classification

- Finalize the corpus annotation and the corpus analysis, release to the community
- Improve the automatic classifier
- Integrate these linguistic information to provide a better semantic control during bilingual pivoting paraphrasing

Thank you for your attention, any question?