Machine Translation error categorization by post-editing automatic analysis

Catherine Kobus, Jean Senellart

Errare Workshop 2013, Ermenonville

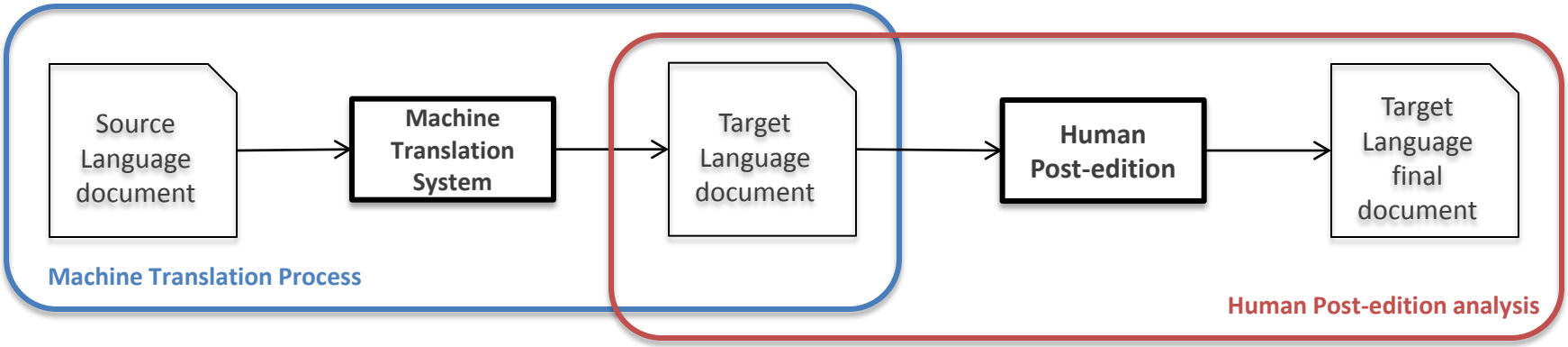
Systran MT system evolution

- « Rule-Based Machine Translation » backbone (till 2006)
  - Engine runs on « linguistic facts » but manually built
  - Deterministic approach (progressively builds the translation)
- Introduction of Statistical Post-Editing (2007-2010)
  - Serial combination of RBMT & SMT [Simard 2007], [Dugast 2007]
  - SPE layer runs on pure text substitution but extracted automatically from bilingual data
  - Approach uses decoding (solution is found and not built)
- Current approach (hybrid)
  - Engine runs on « linguistic facts » but 100% acquired from data
  - Approach uses decoding to select best hypothesis

Post-editing context

- « Highly customized » translation solutions for specific domain/usage
  - Ex: technical documentation, online technical assistance…
- « Corpus-driven » approach
  - SPE layer added
    - ✓ Adaptation to the domain
    - ✓ Better overall translation quality
  - But still need for post-editing
- Big companies (Symantec, Autodesk, Cisco) turn to Post-Editing
  - Cost saving, « time to market »
- Post-editors are professional translators with strict guidelines to follow
  - « Light » post-editing
  - Should not take more time than retranslating from scratch

Qualitative analysis of post-edited data

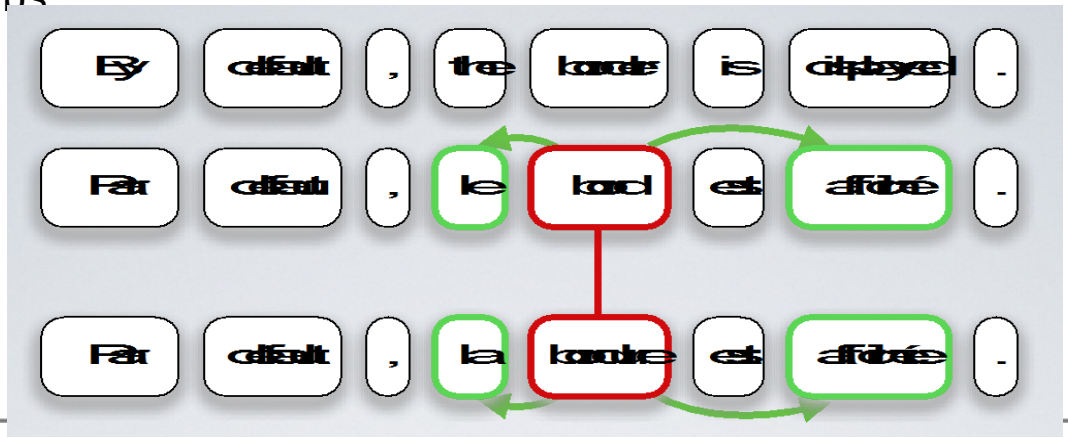


- Qualitative analysis of Human edits
  - Manual analysis of MT outputs for EN->FR
  - Definition of Post-Editing Actions (PEA)
    - ✓ Minimal and logical edits, that make sense linguistically
    - ✓ Opposed to “mechanical” edits (insertion, deletion, substitution, move captured by TER)
  - Goal : better representation of user intent

    => finer evaluation of Machine Translation quality

First Observations

- Post-editing = quite redundant task
  - Many corrections of the same errors, etc
  - Perspective : « on the fly » integration of users' feedback to
    - ✓ reduce post-editing effort
    - ✓ Improve user's experience
- Two « levels » of edits
  - First level : edits related to an error of the MT system
  - Second level: edits induced by a first level edit (i.e. « propagations» of the error)
    - ✓ Ex: in French, gender and number agreement between nouns and adjectives/articles/verbs



PEA typology

- Typology defined for French, based on existing classifications
  - [Font-Llitjòs et al., 2005], [Vilar et al., 2006], [Dugast et al., 2007]
- Four most observed classes
  - **Noun-Phrase (NP)** – related to lexical changes
    - ✓ SRC : the border displays as stripes
    - ✓ TGT : la bordure s'affiche sous forme de **rayures**
    - ✓ PE : la bordure s'affiche sous forme de **bandes**
  - **Verbal-Phrase (VP)** – related to grammatical changes
    - ✓ SRC : connectors can be used
    - ✓ TGT : les connecteurs **peut** être **utilisé**
    - ✓ PE : les conecteurs **peuvent** être **utilisés**
  - **Preposition change**
    - ✓ SRC : snapping to sketches
    - ✓ TGT : accrochage **des** esquisses
    - ✓ PE : accrochage **aux** esquisses
  - **Co-reference change** – through introduction/removal of a pronoun, or change of a definite to possessive determiner
    - ✓ SRC : the distance increases
    - ✓ TGT : **la distance** augmente
    - ✓ PE : **elle** augmente

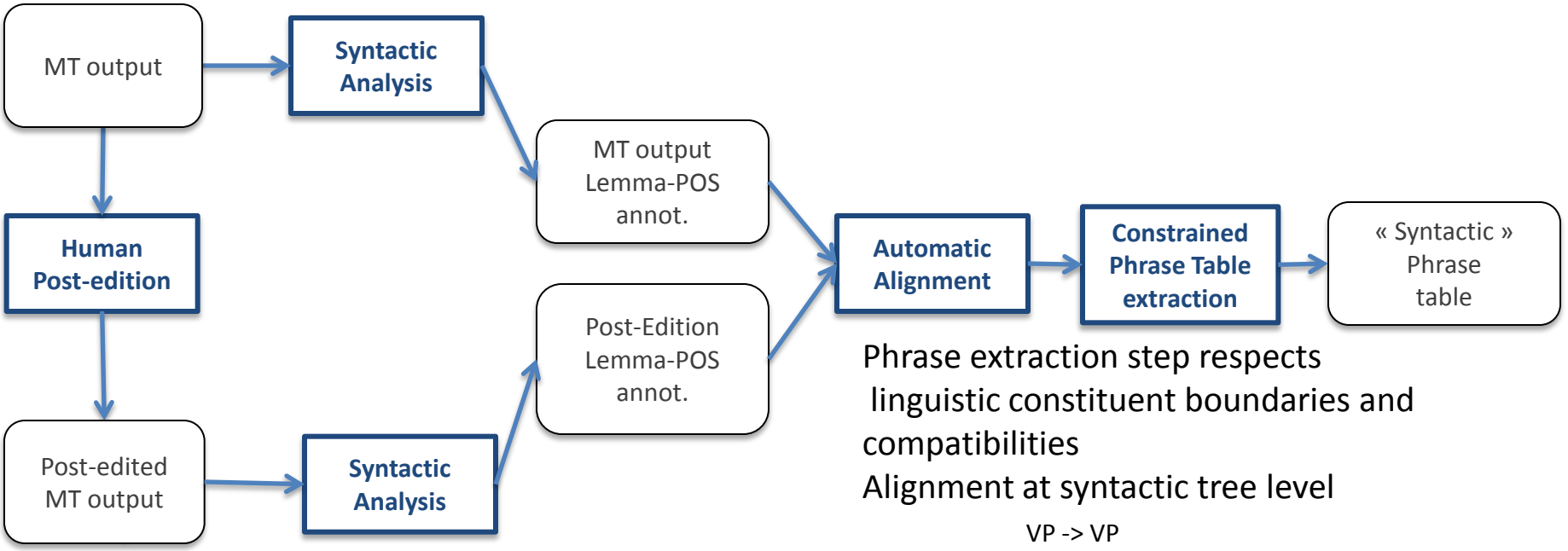
Results

- Main part of the Post-editing effort involves Noun-Phrases (about 90%)
  - Terminological changes (about 60%)
- Top four edits cover around 30% of the total

Class / Sub-class	RBMT system #PEA	RBMT system %PEA
Noun-Phrase (NP)	**74**	**90%**
Determiner choice	1	1.2%
Noun meaning choice	49	59%
Noun number	3	3.6%
Case change	19	23%
Adjective choice	2	2.4%
Verbal-Phrase (VP)	**6**	**7.2%**
Verb agreement change	3	3.6%
Verb meaning choice	3	3.6%
Preposition change	1	1.2%
Co-reference change	2	2.4%
TOTAL	**83**	**100%**

Automatic error categorization

- Need to automate the Post-Editing Actions categorization
  - Extract linguistic rules that can be integrated in our engine
- Linguistically motivated
  - Use of our « syntactic » analysis



Phrase extraction step respects linguistic constituent boundaries and compatibilities

Alignment at syntactic tree level

VP -> VP
NP -> NP
adj -> adj
adj -> PP
adv -> adv
PP -> adv
…

Process

- « Source » and « target » entries that are different in the syntactic phrase table
  - ⇒ Detection and categorization of edits

- Example

  fournir-|-verb:inf ||| atteindre-|-verb:inf

  fournir-|-verb:inf ||| contribuer-|-verb:inf

  fournir-|-verb:inf ||| fournir-|-verb:inf

  fournir-|-verb:inf ||| indiquer-|-verb:inf

  fournir-|-verb:inf ||| offrir-|-verb:inf

  => rules in bold = edits on the word « fournir »

Experiments and results

- Corpus of human post-edited data
  - Technical documentation (computer science domain)
  - About 11400 sentences post-edited
  - Systran MT system : EN -> FR

# words	213907
# errors	34644
# categorized errors	**8621**

- About **25%** of the edits were detected and categorized
  - Crucial information that can be directly incorporated in our engine for Automated Post-Editing
    - ✓ Can be easily converted to bilingual ressources (source -> target)
  - Why not more?
    - ✓ Each step in the process is error-prone
      - Syntactic analysis, alignment, etc
    - ✓ Need more data

Results and analysis

- Most detected edits

génération-|-noun:common	->	build-|-noun:common
normal-|-adj	->	régulier-|-adj
personnage-|-noun:common	->	caractère-|-noun:common
comprendre-|-verb:inf	->	inclure-|-verb:inf
garbage-|-noun:common collection-|-noun:common	->	nettoyage-|-noun:common du-|-prep mémoire-|-noun:common
entrer-|-verb:inf	->	sélectionner-|-verb:inf
indiquer-|-verb:inf	->	montrer-|-verb:inf
accès-|-noun:common concurrentiel-|-adj	->	méthode-|-noun:common Concurrency-|-noun:propernoun
moment-|-noun:common	->	runtime-|-noun:common
du-|-prep execution-|-noun:common		

Results and analysis

- Most detected rules

verb:inf	->	verb:inf	about 20%
noun:common	->	noun:common	about 12%
prep . noun:common	->	prep . noun:common	
adj	->	adj	
pron	->	pron	
noun:common .adj	->	noun:common . noun:propernoun	
prep	->	prep	
noun:common	->	adj	

Conclusions and perspectives

- Qualitative analysis of human edits
  - Definition of Post-Editing Actions
- Automatic detection and categorization
  - Promising results
  - Information can be easily integrated in our engine
- Errors quite redundant and repetitive in Post-Editing context
  - Scores like BLEU do not reflect overall translation quality
  - Need for a global metric to optimize our systems

References

- [Dugast 2007] Dugast L., Senellart J. et Koehn P., Statistical post-editing on systran's rule- based translation system, dans *Proceedings of the Second Workshop on Statistical Machine Translation,* pages 220–223, Association for Computational Linguistics, 2007.
- [Font-Llitjós 2005] Font-Llitjós A., Carbonell J. G. et Lavie A., A framework for interactive and automatic refinement of transfer-based machine translation, dans *European Association of Machine Translation (EAMT) 10th Annual Conference. Budapest, Hungary,* Citeseer, 2005.
- [Simard 2007] Simard M., Ueffing N., Isabelle P. et Kuhn R., Rule-based translation with statistical phrase-based post-editing, dans *Proceedings of the Second Workshop on Statistical Machine Translation,* pages 203–206, Association for Computational Linguistics, 2007.
- [Vilar 2006] Vilar D., Xu J., D'Haro L. F. et Ney H., Error analysis of statistical machine translation output, dans *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC),* Citeseer, 2006.

Thank you!

Catherine Kobus
{kobus, senellart}@systran.fr