



Applying CNL Authoring Support to Improve Machine Translation of Forum Data

Sabine Lehmann Siu Kei Pepe Lo
Ben Gottesman Melanie Siegel
Robert Grabowski Frederik Fouvry
Mayo Kudo

- About ACCEPT
- CNL and MT
- Acrolinx CNL
- User-Generated Content
- Our Approach
- Examples
- Application Scenarios
- Evaluation
- Next Steps



- Enabling machine translation for the emerging community content paradigm.
- Allowing citizens across the EU better access to communities in both commercial and non-profit environments.



UNIVERSITÉ
DE GENÈVE



- Make user-generated content (UGC) easier to read
- Make UGC easier to translate with Machine Translation
(it can't be translated manually)
- UGC is more trusted and more used than company content
- Companies are now trying to make UGC better
 - By “moderating” or “curating” it.

- Fix content before MT: pre-editing rules (CNL)
- Fix content after MT: post-editing rules (CNL)

- CNL and Rules-based MT (RBMT): proven in many cases
 - Symantec with Systran (e.g. thesis: J. Roturier)
 - Thicke, J. Kohl, etc.
- CNL and Statistical MT (SMT): not so clear
 - Working with Moses, Google and Bing
 - Depends on text and training corpus
 - Depends on language pairs

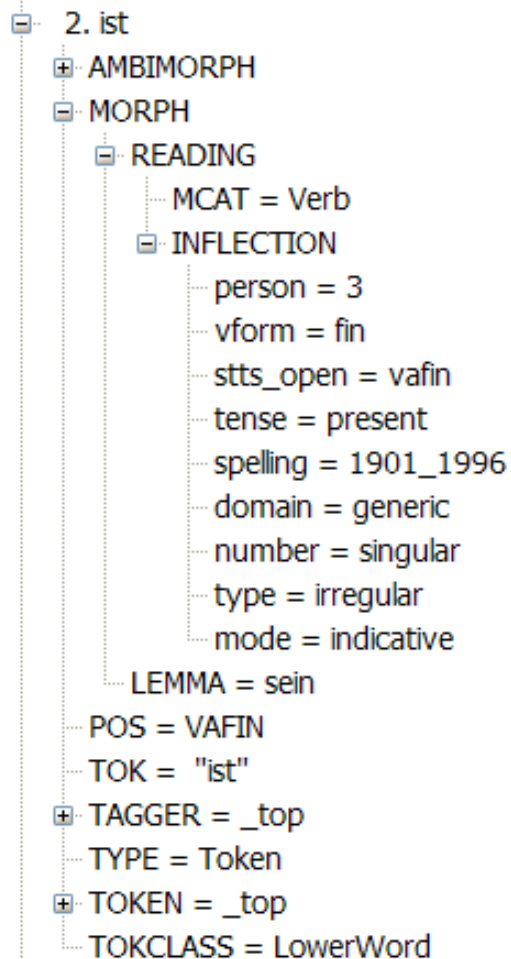
- Acrolinx founded 02.02.02 out of DFKI
- NLP
 - Hybrid system: rule-based with statistical components
 - Multi-level system: Base NLP + Rules Engine
 - Multilingual (EN, DE, FR, JP, ZH, SV, ...)
 - Highly scalable
 - (50k words per second / 10 million words per month)
 - “Looking for errors”
 - More like Information Extraction than Parsing
 - Working with “ill-formed” text

- Tokenizer, Segmentizer
- Morphology
- Decomposition
- POS Tagger, Mecab (for JA and ZH)
- Word Guesser

Additional information

- Terminology (Chunks)
- Gazetteer (Lists of different words)
- Context Information (XML, Word style)

Feature Structure



- “on top” of the basic components
- Acrolinx rule formalism
- Allows user to specify objects based on the information available in the feature structure
- Describing the “locality” of the issue
- Continuous further development of rule formalism based on needs
 - e.g. MT more suggestion possibilities are required

```
//example: a dogs
```

```
TRIGGER(80) == @det_sg^1 [ {@mod|@noun} ]*! @noun_pl^2  
              -> ($det_sg, $noun_pl)  
              -> { mark : $det_sg, $noun_pl; }
```

```
//example: a dogs -> a dog
```

```
SUGGEST(10) == $det_sg [ ]* $noun_pl  
              -> { suggest: $det_sg -> $det_sg, $noun_pl ->  
                  $noun_pl/generateInflections([number="singular"]);  
              }
```

- **Fix content before MT: pre-editing rules (CNL)**
- Fix content after MT: post-editing rules (CNL)
- “Extend” training data

- Informal/spoken language
 - colloquialism
 - truncations
 - Interjections
 - ...
- Use of first person/second person
- Many “questions”
- Ellipses
- In French: lack of accents
- ...

Yes, both the file/app server running Backup Exec ("SERVER01" above) and the SQL server ("SERVER03" above) are running Windows Server 2000. I do not know what AOFO is or where I would check if it's running.

Ahh OK. As a test - for that job that fails - edit the backup job properties and go to the Advanced Open File section.
BTW AOFO = Advanced Open File

Holy crap, Colin, that's exactly what I needed! Thank you. I ran another test job last night with AOFO unchecked and it successfully backed up the PROFXENGAGEMENT database on the SQL server

- avoid parenthetical expressions in the middle of a sentence
- avoid colloquialism
- avoid interjections
- avoid informal language
- avoid complex sentences
- missing end of sentence

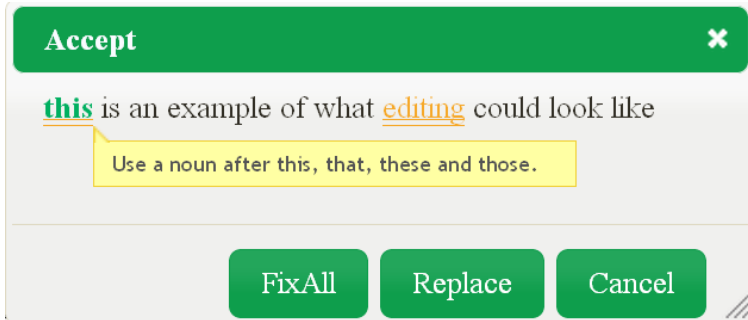
- 512MO ram de **dique** dur, mais **la**, cela a toujours **fonctionner** normalement avant Cela fait 4 jours que le **probleme** est apparu quand des mises **a** jours Windows ont été faites.

- confusion de mots (word confusion)
 - la vs. là
 - ce vs. se
 - a vs. à
- mots simples (simple words)
- évitez questions directes (avoid direct questions)
- évitez le langage familier (avoid informal language)
- évitez moi (avoid specific form of first person pronoun)

- Fix content before MT: pre-editing
- Fix content after MT: post-editing
- **“Extend” training data**

- Not always possible to pre-edit
- Second person typically not in training corpus, but how to get rid of it?
- Use CNL approach (rule formalism) to generate additional training data with second person
vous cliquez -> tu cliques

- Interactive (Plug-ins to forums)



- Automatic (also for training data)

- Automatic pre-editing replaces suggestion automatically
instalation -> installation
- generally very difficult because precision needs to be very high
- tests done with autoApplyClient

- automatically replaces marked sections of text with the top-ranked improvement suggestion given by Acrolinx
- Use Cases
 - automatic pre-editing
 - evaluation

- idea to work with sequential rule sets
 - some rules need to apply before others
 - order rules into different rule sets wrt their order in which they have to apply
- EN: currently 6 rule sets
- FR: tests started last week!

- I am trying to setup that feature, but it doesnot **work** **What** am I missing?

----- segmentation rules -----

- I am trying to setup that feature, but it doesnot **work.** **What** am I missing?

- I am trying to setup that feature, but it **doesnot** work. What am I missing?

----- spelling -----

- I am trying to setup that feature, but it **does not** work. What am I missing?

- I am trying to **setup** that feature, but it does not work. What am I missing?

----- **specific grammar rules** -----

- I am trying to **set up** that feature, but it does not work. What am I missing?

- Automatically apply Acrolinx rules
- Evaluate with respect to
 - BLEU (Bilingual Evaluation Understudy)
 - GTM (General Text Matcher)
 - TER (Translation Error Rate)

- MT is improved
 - Automatic correction correlates with human evaluation

	Human		GTM		BLEU		TER	
	+	-	+	-	+	-	+	-
Use relative pronouns such as that and which	33%	4%	26%	19%	26%	7%	15%	15%
Confusion between noun and adjective	23%	8%	46%	0%	46%	0%	38%	8%
Avoid contractions	27%	12%	31%	12%	31%	8%	27%	19%

- Focus more on corpus
 - unknown word in the training data
 - check frequency of rules in the training data to infer whether rule is relevant
- Post-editing for SMT
- More evaluation

Thank You!

Sabine Lehmann

sabine.lehmann@acrolinx.com