Handling Idioms in Symbolic Multilingual Natural Language Generation

MICHAELLE DUBÉ, FRANÇOIS LAREAU
OLST, Université de Montréal
IDIOMS IN RULED BASES NLG

- Non-compositional MWE
  - Single unit of meaning
  - Syntactic structure
- Syntactic cohesion
- Form flexibility

Introduction

'sick'

SICK

'UNDER THE WEATHER'
1. **Donner encore sa langue au chat**
   ‘give up guessing again’, lit. ‘give again one’s tongue to the cat’

2. **Take it with a big grain of salt.**

3. **You must take whatever he says with a grain of salt.**

4. **He’s just pulling your leg.**

5. **‘LIETUVOS APELIACINIS TEISMAS’**
   ‘court of appeals of Lithuania’ lit. ‘appellate court of Lithuania’

6. **It was a pretty big bullet to bite**

- modifier
- complement
- inflection
Overview

Linguistic Realizers
Lexical Data
Implementation
Evaluation
Linguistic Realizers

Lexical Choice

SimpleNLG
RealPro
JSReal
ATML3

Output Flexibility

KPML
Open CCG
FLAUBERT
FUF/SURGE

Semantic Input

MARQUIS
FORGe
GenDR

Syntactic Input
Lexical Data

- French Lexical Network (LN_{fr})
- Grammatical classification
- Linear syntactic patterns
- Sequences of part of speech (POS)

| Linguistic patterns | Example                | #   | %   |
|---------------------|------------------------|-----|-----|
| N Prep N            | TÊTE DE MULE           | 409 | 14% |
| N Adj               | TERRE FERME            | 377 | 13% |
| Prep N              | DE JUSTESSE            | 222 |  8% |
| N Prep.circ N       | CORPS À CORPS          | 138 |  5% |
| Adj N               | JOLI COEUR             | 100 |  4% |
| Prep Det N          | DANS LE VENT           |  98 |  3% |
| V Det N             | LEVER LE PIED          |  97 |  3% |
| N Prep N            | ART DE LA TABLE        |  97 |  3% |
| Others              | …                      | 1417| 49% |
| **Total**           |                        | **2919** | **100%** |
Implementation

- Generic deep realizer (GenDR)
- Resource sharing
- MATE = graph transducer
- Meaning-Text theory (MTT)
- Semantics-syntax interface

Diagram:

```
  'courtiser'
    ▼
   ▼
  'Sam'  'Alex'  F subtitle
```

```
  SAM
  ▼
  LE
  ▼
  ALEX
```

SemR  SSyntR
Template Lexicalization Rules

```
DSyntR

JOINDRE LES
DEUX BOUTS

⇒

JOINDRE

⇒

bouts

⇒

les
deux

SSyntR
```
\[\text{JOINDRE LES DEUX BOUTS} \quad \text{ENFONCER UNE PORTE OUVERTE} \quad \text{DANS DE BEAUX DRAPS}\]

\[
\begin{align*}
\text{V Det Num N} & \quad \Rightarrow \quad \text{V Det N Adj} \\
\text{Prep Det Adj N} & \quad \Rightarrow \quad \text{Generic}
\end{align*}
\]
Generic Tree Patterns

| Partition | # tree |
|-----------|--------|
| 4 nodes   |        |
| 3         | 2      |
| 1 + 2     | 1      |
| 1 + 1 + 1 | 1      |
|           | =4     |

4.01

\[
\begin{align*}
\text{\(w_0\)} & \quad \text{\(w_1\)} \\
\downarrow & \quad \downarrow \\
\text{\(w_2\)} & \quad \text{\(w_0\)} \\
\downarrow & \quad \downarrow \\
\text{\(w_3\)} & \quad \text{\(w_1\)} \\
\end{align*}
\]

4.02

\[
\begin{align*}
\text{\(w_0\)} & \quad \text{\(w_1\)} \\
\downarrow & \quad \downarrow \\
\text{\(w_2\)} & \quad \text{\(w_3\)} \\
\end{align*}
\]

4.03

\[
\begin{align*}
\text{\(w_0\)} & \quad \text{\(w_1\)} \\
\downarrow & \quad \downarrow \\
\text{\(w_2\)} & \quad \text{\(w_3\)} \\
\end{align*}
\]

4.04

\[
\begin{align*}
\text{\(w_0\)} & \quad \text{\(w_1\)} \\
\downarrow & \quad \downarrow \\
\text{\(w_2\)} & \quad \text{\(w_3\)} \\
\end{align*}
\]
Mapping Patterns

Input

\[ w_0 \]
\[ \downarrow \]
\[ w_1 \]
\[ w_2 \]
\[ \downarrow \]
\[ w_0 \]
\[ \downarrow \]
\[ w_2 \]

\[ w_3 \]
\[ \downarrow \]
\[ w_3 \]
\[ \downarrow \]
\[ w_1 \]

V Det Num N

Mapping

\[ joindre \]
\[ \downarrow \]
\[ V \]
\[ \downarrow \]
\[ Det \]
\[ \downarrow \]
\[ deux \]
\[ \downarrow \]
\[ Num \]
\[ \downarrow \]
\[ bouts \]

Output

\[ joindre \]
\[ \downarrow \]
\[ bouts \]
\[ \downarrow \]
\[ les \]
\[ \downarrow \]
\[ deux \]
5. Evaluation

5.1 Coverage
5.2 Precision
Results: Coverage

| Pattern Type          | Coverage (%) | # Trees | # Idioms | Coverage Limit |
|-----------------------|--------------|---------|----------|----------------|
| Idioms                | 97.5%        | 2846    | 2919     |                |
| Linguistic Patterns   | 87.9%        | 452     | 514      |                |
| Generic Patterns      | 80.6%        | 29      | 36       |                |

Graph showing the number of trees, idioms, and coverage limit over iterations.
97.8% OVERALL PRECISION

| Sample   | n  | Judge A | Judge B | k   |
|----------|----|---------|---------|-----|
| Sample 1 | 100| 98      |         |     |
| Sample 2 | 100|         | 97      |     |
| Sample 3 | 300| 295     | 294     | 0.91|

PROBLEMS / SOLUTIONS

- Pattern mapping
  - Vague linguistic pattern
  - Node inflection
- Data collection (inflection)
- LNfr precision
  - Government patterns & coreferences
6. Conclusion

Handling MWEs in MNLG
- Language independent
- Linguistic patterns
- Generic patterns

Future work
- Graph transducer
  - GREW
- Extend to English and Russian
  - LN\textsubscript{en} LN\textsubscript{ru}
2846 French Idioms

452 Linguistic Patterns

36 Generic Patterns

Linguistic Patterns

ENGLISH MWEs

RUSSIAN MWEs
Thank you

Michaëlle Dubé
michaëlle.dube@umontreal.ca
www.olst.ling.umontreal.ca