Token-based MWE Identification Strategies in the mwetoolkit

Silvio Ricardo Cordeiro Carlos Ramisch Aline Villavicencio

Aix Marseille Université, CNRS, LIF UMR 7279, 13288, Marseille, France

Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

The mwetoolkit in a nutshell

- ▷ The toolkit: General-purpose collection of tools for the automatic **identification** and processing of **MWEs**;
- ▷ Previous functionalities: multilevel regex-like MWE **extraction**, efficient n-gram **counting**, **filtering**, **measuring** ...
- ▷ Motivation for this work: **lack** of **expressive patterns** limited what MWEs could be found, and the restricted alternatives for **corpus annotation** also lacked the ability to project back into source.

New: Expressive extraction patterns

```
<pat>  <!-- Finding past forms of "wake (up)" -->
  <w lemma="wake" pos="Verb">
     <neg surface="wake"/> <neg surface="wakes"/>
  </w>
  <pat repeat="?">
     <w surface="up"/>
  </pat>
</pat>
```

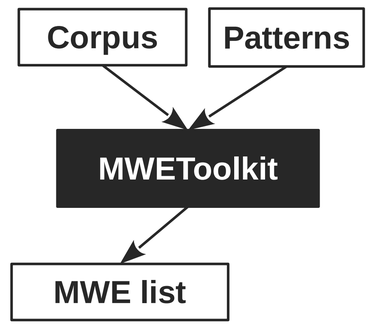
New: MWE in-corpus annotation

MWE extraction & annotation in corpora:

- ↪ Different gapping possibilities: contiguous, gappy;
- ↪ Different match distances: shortest, longest, all;
- ↪ Different match modes: overlapping or not;
- ↪ Annotation projecting back extracted MWEs;
- ↪ Annotation based on external lists of MWEs.

Extracting the pattern above from a POS-tagged corpus

In the **wake** of the war ⇒ ~~Nothing~~ (POS ≠ "Verb")
against the **woken** dragon ⇒ woken
a hero is **wakened up** again. ⇒ wakened up
The **woken** dragon is defeated ⇒ woken
but other dragons still **remain**! ⇒ ~~Nothing~~ (lemma ≠ "wake")
Do not **wake** them **up** ... ⇒ ~~Nothing~~ (surface = "wake")



Annotating a corpus with different match distances

Pattern **Verb (Word*) Particle**

- ○ Shortest: I have **picked** it **up** and **put** it **down**. ✓
- ○ Longest: I have **picked** it up and put it **down**. ✗

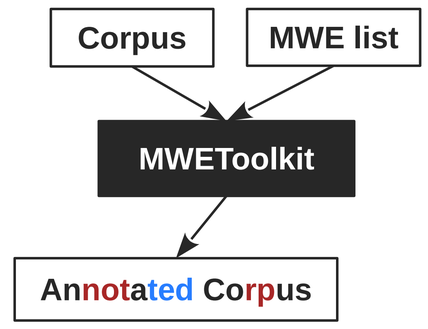
Pattern **Noun Noun$^+$**

- ○ Shortest: The **science fiction** writers are on strike. ✗
- ○ Longest: The **science fiction writers** are on strike. ✓

Annotating a corpus with gaps and overlapping MWEs

Using both patterns above

- ○ S+S: You **threw** all those **lab rat** tissue samples **out** without thinking! ✗
- ○ L+L: You **threw** all those **lab rat tissue samples** out **without** thinking! ✗
- ○ S+L: You **threw** all those **lab rat tissue samples out** without thinking! ✓



Future work

- ○ Consider surrounding context when annotating:
  1. The test was a **piece of cake**. ⇒ MWE
  2. I just ate a big **piece of cake**. ⇒ Non-MWE
- ○ Use sequence models to disambiguate MWE occurrences in the text (so that only idiosyncratic cases are annotated).
- ○ Define measures and implement tools to evaluate MWE annotation quality in corpora.

Financial support from the CAMELEON project (CAPES-COFECUB 707-11) and the AIM-WEST project.