Token-based identification of Multiword Expressions

Silvio Ricardo Cordeiro

Advisor: Aline Villavicencio
Co-advisor: Carlos Ramisch

October 24, 2014

Outline

1. Introduction
2. Token-based identification
3. Conclusions

Outline

1 Introduction

2 Token-based identification

3 Conclusions

Multiword Expressions (MWEs)

- Group of lexemes in a syntactic unit;
- Semantics cross single-lexeme boundaries;
- Idiosyncratic **non-compositional** interpretation;

Examples

- "To kick the bucket";
- "To look ⟨something⟩ up";
- "Silver bullet";
- "By and large".

Research Motivation

- As **frequent** as single words [Jackendoff, 1997];
- **Requirement** for large-scale NLP applications [Sag et al., 2002];
- State-of-the-art is **lacking** [Rayson et al., 2010; Schneider et al., 2014].

General objective

- Have a pipeline that extends from the detection to the parallel annotation and paraphrasing of Multiword Expressions.
- Adapt applications to use this information.

Applications

- Machine translation:
    - “He kicked the bucket” → “Ele bateu as botas”;
    - “The police car” → “A viatura policial”.
- Text simplification:
    - “He kicked the bucket” → “He died”;
    - “The malaria mosquito” → “The mosquito that transmits malaria”.

Targeted objectives

- Generic token-based pattern description;
- Pattern-based MWE identification & extraction;
- Measure MWE identification;
- Annotate MWEs in parallel corpora;
- Apply techniques:
    - Machine Translation;
    - Text Simplification.

Outline

1 Introduction

2 Token-based identification

3 Conclusions

MWE extraction and measure

1. Annotating a corpus based on MWE patterns;
    1. Control pattern identification length;
    2. Overlapping of MWEs matching a pattern;
    3. Improve $\mathcal{O}(n^2)$ annotation scheme of Kulkarni [2011];
2. Representation of attribute-negation pattern;
3. Develop a linear-scale MWE identification measure;
    - Compare with exact-match and link-based identification measures [Schneider et al., 2014];

Task

Detect **MWE patterns** in a list of **tokens** [Ramisch et al., 2010].

Pattern

`N N N*`

Tokens

A mouse liver cell line was derived from MMH-D3 cells after 40 serial passages under suboptimal conditions [...]

Detected MWE

mouse liver cell line

Problem

- Pattern: Verb (Whatever*) Particle.
- Tokens: I have picked it up and put it down.
- Annotation: I have picked it up and put it down.

Solution: Match lengths & overlap control

- Long: I have picked it up and put it down.
- Short: I have picked it up and put it down.
- Overlap: I have picked it up and put it down.

Problem

We may have a list of MWEs and non-annotated tokens.

Solution

Annotate tokens from list of **MWEs** [Kulkarni et al., 2011];

MWEs

derive from, liver cell line, mouse liver, put up, serial passages . . .

Annotated Tokens

A mouse liver cell line was derived from MMH-D3 cells after 40 serial passages under suboptimal conditions [...]

MWE extraction and measure

1. ~~Annotating a corpus based on MWE patterns~~;
   1. ~~Control pattern identification length~~;
   2. ~~Overlapping pattern identification~~;
   3. ~~Improve $\mathcal{O}(n^2)$ annotation scheme of Kulkarni [2011]~~;
2. Representation of attribute-negation pattern;
3. Develop a linear-scale MWE identification measure;
   - Compare with exact-match and link-based identification measures [Schneider et al., 2014].

Old “Negation” pattern

- `<w lemma="walk" pos="V" neg="lemma:pos"/>`

Problem

- Cannot represent some negations...
- `<w lemma="smart" pos="V" pos=N" neg="pos"/>`

New “Negation” pattern

-
  ```
  <w lemma="smart">
       <neg pos="V"/> <neg pos="N"/>  </w>
  ```
- Any boolean expression in DNF!

MWE extraction and measure

1. ~~Annotating a corpus based on MWE patterns~~;
    1. ~~Control pattern identification length~~;
    2. ~~Overlapping pattern identification~~;
    3. ~~Improve $\mathcal{O}(n^2)$ annotation scheme of Kulkarni [2011]~~;
2. ~~Representation of attribute-negation pattern~~;
3. Develop a linear-scale MWE identification measure;
    - Compare with exact-match and link-based identification measures [Schneider et al., 2014].

Reference vs Prediction

- Reference corpus:
  My wife has taken her car in for a routine oil change.
- Predicted annotation:
  My wife has taken her car in for a routine oil change.

Measuring matches (precision, recall)

- Traditional: $^0/_1$, $^0/_1$.
- Link-based: $^1/_2$, $^1/_1$ [Schneider et al., 2014].
  - Predicted links: $\{(taken, in), (in, for)\}$.

Our goal

- Some MWEs are more predictable / compositional.
- Predicting look up is more important than come back;
- Therefore: Measure in a continuum.

Outline

1. Introduction
2. Token-based identification
3. Conclusions

Final thoughts

- MWE semantics is hard;
- MWE extraction & annotation is lacking;
- Patterns can express many criteria;
- Better measures → understand identification mistakes.

Future work

1. Annotate MWEs in parallel corpora;
   1. With corpus alignment information;
   2. With pattern-based constraints;
   3. With paired (source-target) patterns;
2. Annotation-based MWE paraphrasing;
   1. Lexical semantic segmentation [Schneider et al., 2014];
3. Use this information in applications.

References

Jackendoff, R. (1997). Twistin'the night away. *Language*, pages 534–559.

Kulkarni, N. and Finlayson, M. A. (2011). jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124. Association for Computational Linguistics.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). Multiword expressions in the wild?: the mwetoolkit comes in handy. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 57–60. Association for Computational Linguistics.

Rayson, P., Piao, S., Sharoff, S., Evert, S., and Moirón, B. V. (2010). Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, 44(1):1–5.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Schneider, N., Danchik, S. O. N. K. E., Noah, M. T. M. H. C., and Smith, A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. *Proc. of LREC. Reykjavík, Iceland*.

Token-based identification of Multiword Expressions

Silvio Ricardo Cordeiro

Advisor: Aline Villavicencio
Co-advisor: Carlos Ramisch

October 24, 2014