Without lexicons, MWE identification will never fly: A position statement

Agata Savary[1], Silvio Ricardo Cordeiro[2], Carlos Ramisch[3]

[1]Université de Tours, [2]Paris-Diderot University, [3]Aix Marseille Université, **France**

MWE-WN@ACL, Florence, 2 Aug 2019

Multiword expression identification (MWEI)

Task definition

Point at occurrences of MWEs in running text

Distinguish MWEs from regular combinations:

$\Longrightarrow$ *to **take pains*** 'to try hard' vs. *to take gloves*

State of the art

- 2 editions of the PARSEME shared task on automatic identification of verbal MWEs (2017 & 2018)
- MWEI is more challenging than related tasks (e.g. NER)

Position statement

- The difficulties of MWEI lie in the very **nature of MWEs**
- MWEI should systematically be coupled MWE discovery via **NLP-applicable syntactic lexicons of MWEs**

Outline

Outline

Multiword expressions (MWEs)

Diversity of multiword expressions

- compounds: ***to and fro***, ***crystal clear***, *a* ***slam dunk*** 'an easily achieved victory'
- verbal idioms: *to* ***take pains*** 'to try hard'
- light-verb constructions: *to* ***pay*** *a* ***visit***
- verb-particle constructions: *to* ***take off***
- institutionalized phrases: ***traffic light***
- multiword terms: ***neural network***
- multiword named entities: ***Federal Bureau of Investigation***

MWE dichotomy

Sublanguage MWEs (SL-MWEs)

- multiword **named entities** (NEs) and **multiword terms**
- coined by sublanguage **experts** via dedicated **nomenclature instruments** (e.g. scientific publications, naming committees)

General language MWEs (GL-MWEs)

- coined by much larger communities of speakers via informal processes

MWE properties I

Proliferation speed (P_{prolif})

- SL-MWEs **strongly proliferate**
- GL-MWE take **longer to establish** in a language

Nature of discrepancies (P_{discr})

- SL-MWEs - peculiarities at the level of **tokens** (individual occurrences)
  - multiword NEs - capitalization, trigger words (*Bureau*, *river*, *Mr.*)
  - multiword terms - components are rarer in general language (*neural*)
- GL-MWE - mostly regular at the level of tokens, idiosyncratic at the level of **types** (sets of surface realizations of an MWE)
  - *to **take pains*** 'to try hard', *#to take the pain*
  - *to take gloves*, *to take the glove*

MWE properties II

Component similarity (P_{sim})

- SL-MWEs - strong surface/semantic similarity of components
  - Modification of previous terms:
    - ***neural network***, ***neural net***, ***recurrent neural network***, ***neural network pushdown automata***
  - Lexical replacement within a given semantic class:
    - ***American/Brazilian/French/Ethiopian Red Cross***, ***Nigerian Red Cross Society***, ***Iranian/Iraki Red Crescent Society***, ***Saudi Red Crescent Authority***
- GL-MWE - moderate similarity of components
  - LVCs - few frequent light verbs, nouns always predicative; but: the same verbs are also highly frequent in regular constructions:
    - ***make a decision***, ***pay a visit*** vs. *to make bread*
  - IRVs - verb always governs the RCLI, RCLI hardly inflects; but: synonymous verbs are not necessarily inherently reflexive:
    - PL ***znaleźć się*** 'find oneself' vs. PL ***wyszukać się** 'find oneself'
  - VIDs - dissimilar to each other but similar to regular constructions
    - *to* ***take pains*** 'to try hard' vs. *to take aches*

Outline

Identification of sublanguage MWEs (NEs, terms) I

CoNLL 2002 and 2003 shared task on **named entity recognition**

Language	annotated NEs	Best 2002/2003	Best 2018
German	20K	0.71	0.78
Dutch	13K	0.74	0.85
Spanish	18K	0.77	0.85
English	35K	0.86	0.90

2002 and 2003 results

- Machine learning: HMM, decision tree, MaxEnt, CRF, SVM
- Heavy use of external lexicons (gazeteers)

2018 results

- Up-to-date results by Yadav and Bethard [2018]
- Deep neural networks, no lexicon lookup

Identification of sublanguage MWEs (NEs, terms) II

Term identification

- Several domain-specific benchmarks [Campos et al., 2012]
  - F1=0.81 on disorder names
  - F1=0.85 on chemical names
  - F1=0.88 on gene/protein names

Discussion

- Morphology affects results (F1=0.71–0.77 in Polish NER)
- Single-word and multiword entities considered
  $\Longrightarrow$ But multiword NEs and terms are very frequent
- Machine learning, reasonably good F1 scores for ∼20 years

Identification of general-language MWEs I

Focus on PARSEME 1.1 shared task

- 19 languages, verbal MWEs
- Best systems: average from F1=0.5 to F1=0.58

Overview of "largest" languages:

	BG	**FR**	**PL**	**PT**	**RO**	**TR**
#verbal MWEs	6.7K	5.7K	5.2K	5.5K	5.9K	7.1K
unseen ratio	.33	.50	.28	.28	.05	.75
Best non-NN F1	.63	.56	.67	.62	.83	.45
Best NN F1	.66	.61	.64	.68	.87	.59

Identification of general-language MWEs II

Discussion

- MWE scores do not exceed 0.68
  $\Longrightarrow$ except for RO, which has a low unseen ratio
- Hard to compare results on SL-MWEs and GL-MWEs
  - Categories included in NER
  - Single- and multiword entities mixed
  - GL-MWE corpora are much smaller
- Still, MWEI seems to be a particularly hard problem

Challenges of unseen data

- Best open (SHOMA) and closed (TRAVERSAL) track systems
- Phenomenon-specific measures: seen vs. unseen

		BG	PL	PT
TRAVERSAL	seen	.76	.85	.78
	unseen	.13	.17	**.20**
SHOMA	seen	.78	.82	.87
	unseen	**.31**	.18	**.31**

- Better generalization for unseen LVCs and IRVs (P_{sim})
- Very low when compared to unseen SL-MWEs
  $\Longrightarrow$ F1=0.81 to F1=0.94 on unseen NEs [Augenstein et al., 2017]
- Unseen GL-MWEs seem harder than unseen SL-MWEs
  - P_{discr} and P_{sim} differences
  - Machine learning can more easily take unseen SL-MWEs into account

Potential progress in seen GL-MWEs

- GL-MWEs have low ambiguity (P_{ambig})
  - F1=0.88 for French with simple baseline [Pasquer et al., 2018]
  - Similar approach ranked second in DiMSUM task [Cordeiro et al., 2016]

		BG	PL	PT
TRAVERSAL	identical to train	.85	.92	.87
	variants of train	.55	.80	.72
SHOMA	identical to train	.89	.95	.93
	variants of train	.52	.71	.81

- Room for improvement in discontinuity representation
  $\Longrightarrow$ neural nets with self-attention mechanism [Rohanian et al., 2019]
- **Syntactic MWE lexicons** can cover variants

Outline

MWE lexicons

- Lexicographic tradition
- Encoding formalisms [Gross, 1986, Mel'čuk et al., 1988, Pausé, 2018]
- Partial NLP applicability [Constant and Tolone, 2010, Lareau et al., 2012]
- Losnegaard et al. [2016] present a survey on MWE lexicons

3 important aspects

1. account of the morpho-syntactic structure (variants)
2. lexicon-corpus coupling
3. coverage (number of entries)

1. Morpho-syntactic structure I

Simple

- Raw list
- Raw list + some variations [Steinberger et al., 2011]

More elaborate

- Finite-state technology: POS and morphology of components
  Karttunen et al. [1992], Breidt et al. [1996], Oflazer et al. [2004],. . .
- Continuous MWEs, local morphosyntactic phenomena
- Intentional format (rules) vs. extensional format (rule application)
- No account of deeper syntax, open slots
  $\Longrightarrow$ not ideal for many verbal MWEs

1. Morpho-syntactic structure II

Lexicons not focusing on MWEs

- Theory-neutral approaches [Grégoire, 2010, Przepiórkowski et al., 2017, McShane et al., 2015]

  $\Longrightarrow$ implicit regular grammar – lexicon explicitly encodes irregularities
- Approaches specific to syntactic theories: HPSG, LFG, TAG, etc.

2. Lexicon-corpus coupling

- Fully aligned lexicons: PDT-Vallex [Urešová, 2012], SemLex [Bejček and Straňák, 2010]
- Partly aligned lexicons (corpus examples): Walenty [Przepiórkowski et al., 2014]
- Lexicon entries extracted from raw corpora: DUELME [Grégoire, 2010]

3. Number of entries

- **Great variability**
  $\Longrightarrow$ from a few dozen to tens of thousands of entries
- **Coverage**
  $\Longrightarrow$ often inversely proportional to the richness and precision

MWE lexicons in MWEI

Sequence tagging methods (CRF, perceptron, etc.)

- Constant et al. [2013] and Schneider et al. [2014] show that handcrafted lexicons provide important features for high-coverage MWEI
- Riedl and Biemann [2016] show that discovered lexicons help MWEI

PARSEME shared task

- Only one (rule-based) system uses lexicons [Nerima et al., 2017]
- Maybe because of focus on multilingualism?
  - No unified lexicon format
  - High variability of verbal MWEs requires complex lexical encoding
  - Integration with machine learning methods is not straightforward

Outline

Towards syntactic MWE lexicons

The situation

- MWEI systems must be able to generalize over unseen data (because of P_{zipf})
- MWEI systems must take variability into account to handle seen data

The question

- How to maximize the amount of the seen data at reasonable cost?

An idea

-

Roadmap

Bibliography I

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. Generalisation in named entity recognition. Comput. Speech Lang., 44(C):61–83, July 2017. ISSN 0885-2308. doi: 10.1016/j.csl.2017.01.012. URL https://doi.org/10.1016/j.csl.2017.01.012.

Eduard Bejček and Pavel Straňák. Annotation of multiword expressions in the Prague dependency treebank. Language Resources and Evaluation, 44(1–2):7–21, 2010.

Elisabeth Breidt, Frédérique Segond, and Guiseppe Valetto. Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In Proceedings of COLING-96, Copenhagen, pages 1036–1040, 1996.

David Campos, Sérgio Matos, and José Luís Oliveira. Biomedical named entity recognition: A survey of machine-learning tools. In Shigeaki Sakurai, editor, Theory and Applications for Advanced Text Mining, chapter 8. IntechOpen, Rijeka, 2012. doi: 10.5772/51066. URL https://doi.org/10.5772/51066.

Matthieu Constant and Elsa Tolone. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In Michele De Gioia, editor, Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie, volume 1 of Lingue d'Europa e del Mediterraneo, Grammatica comparata, pages 79–93. Aracne, April 2010. URL http://www.aracneeditrice.it/aracneweb/index.php/catalogo/9788854831667-detail.html. ISBN 978-88-548-3166-7.

Matthieu Constant, Joseph Le Roux, and Anthony Sigogne. Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. TSLP Special Issue on MWEs: from theory to practice and use, part 2 (TSLP), 10(3), 2013. ISSN 1550-4875.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. UFRGS&LIF at SemEval-2016 task 10: Rule-based MWE identification and predominant-supersense tagging. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 910–917, San Diego, California, USA, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1140. URL https://www.aclweb.org/anthology/S16-1140.

Bibliography II

Nicole Grégoire. DuELME: a Dutch electronic lexicon of multiword expressions. Language Resources and Evaluation, 44(1-2), 2010.

Maurice Gross. Lexicon-grammar: The Representation of Compound Words. In Proceedings of the 11th Coference on Computational Linguistics, COLING '86, pages 1–6, Stroudsburg, PA, USA, 1986. Association for Computational Linguistics. doi: 10.3115/991365.991367. URL http://dx.doi.org/10.3115/991365.991367.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. Extension of Zipf's law to words and phrases. In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02, pages 1–6, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1072228.1072345. URL https://doi.org/10.3115/1072228.1072345.

Lauri Karttunen, Ronald M. Kaplan, and Annie Zaenen. Two-Level Morphology with Composition. In Proceedings of COLING-92, Nantes, pages 141–148, 1992.

François Lareau, Mark Dras, Benjamin Boerschinger, and Myfany Turpin. Implementing lexical functions in xle. 06 2012. doi: 10.13140/2.1.2869.9201.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. Parseme survey on mwe resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Marjorie McShane, Sergei Nirenburg, and Stephen Beale. The Ontological Semantic treatment of multiword expressions. Lingvisticæ Investigationes, 38(1):73–110, 2015.

Bibliography III

Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, and Suzanne Mantha.
Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques, volume II of Recherches lexico-sémantiques. Presses de l'Univ. de Montréal, 1988. URL http://books.google.fr/books?id=zwObmgEACAAJ.

Luka Nerima, Vasiliki Foufi, and Eric Wehrli. Parsing and MWE detection: Fips at the PARSEME shared task. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 54–59, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1706. URL https://www.aclweb.org/anthology/W17-1706.

Kemal Oflazer, Özlem Çetonoğlu, and Bilge Say. Integrating Morphology with Multi-word Expression Processing in Turkish. In Second ACL Workshop on Multiword Expressions, July 2004, pages 64–71, 2004.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. If you've seen some, you've seen them all: Identifying variants of multiword expressions. In Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics. The COLING 2018 Organizing Committee, 2018.

Marie-Sophie Pausé. Modelling french idioms in a lexical network. Studi e Saggi Linguistici, 55(2):137–155, 2018. URL https://www.studiesaggilinguistici.it/index.php/ssl/article/view/210.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. Extended phraseological information in a valence dictionary for NLP applications. In Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), pages 83–91, Dublin, Ireland, 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/siglex.html#2014_0.

Adam Przepiórkowski, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. International Journal of Lexicography, 30(1):1–38, 2017. URL http://ijl.oxfordjournals.org/content/early/2016/02/22/ijl.ecv048.abstract?keytype=ref&ijkey=jWNJn7Cxf7WJRhD.

Bibliography IV

Martin Riedl and Chris Biemann. Impact of MWE resources on multiword recognition. In Proceedings of the 12th Workshop on Multiword Expressions, (MWE 2016), Berlin, Germany, August 2016. URL http://aclweb.org/anthology/W/W16/W16-1816.pdf.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. Bridging the gap: Attending to discontinuity in identification of multiword expressions. CoRR, abs/1902.10667, 2019. URL http://arxiv.org/abs/1902.10667.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. Zipf's law holds for phrases, not words. Scientific Reports, 5, 2015. doi: 10.1038/srep12209. URL https://www.nature.com/articles/srep12209.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa I nurrieta, and Voula Giouli. Literal occurrences of multiword expressions: Rare birds that cause a stir. The Prague Bulletin of Mathematical Linguistics, 112:5–54, April 2019. ISSN 0032-6585. doi: 10.2478/pralin-2019-0001. URL https://ufal.mff.cuni.cz/pbml/112/art-savary-et-al.pdf.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. Transactions of the ACL, 2:193–206, 2014.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. JRC-NAMES: A freely available, highly multilingual named entity resource. In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, pages 104–110, Hissar, Bulgaria, September 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/R11-1015.

Zdeňka Urešová. Building the PDT-Vallex valency lexicon. In On-line Proceedings of the fifth Corpus Linguistics Conference, University of Liverpool, 2012.

Jakub Waszczuk, Agata Savary, and Yannick Parmentier. Promoting multiword expressions in A* TAG parsing. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 429–439, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1042.

Bibliography V

Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2145–2158, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1182.