A Neural Graph-based Approach to VMWE Identification

Jakub Waszczuk & Rafael Ehren & Regina Stodden & Laura Kallmeyer

Heinrich Heine University, Düsseldorf, Germany

Method

- **Assumption**: VMWEs local in dependency structures [1,2,3,4]
- **Orchestration**: dependency parsing $\Longrightarrow$ VMWE identification
- **Reduction**: VMWE identification $\Longrightarrow$ dependency tree labeling [4,5]
- **Arc-factored**: each arc separately scored as to its affinity of being a VMWE
- **Neural ingredient**: scoring performed using a MLP (and derivatives)

Basic encoding

- **Labeling**: function $\ell_E\colon E \to \mathbb{B}$ defined over the dependency arcs $E \subset V \times V$
- **Encoding**: $\ell_E(v, w) \coloneqq 1$ iff both v and w belong to a single VMWE occurrence
- **Decoding**: adjacent 1-labeled arcs assumed to form a single VMWE occurrence
- **No support** for single-token, disconnected, or overlapping VMWE occurrences

Extended encoding

- **Labeling**: arc and node labeling functions $\ell_E\colon E \to \mathbb{B}$ and $\ell_V\colon V \to \mathbb{B}$
- **Limitation:** inability to represent overlapping VMWE occurrences

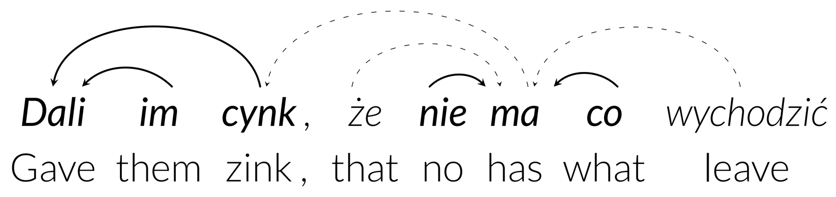

Figure 1: Extended encoding applied to two Polish idioms, *dać komuś cynk* \`give someone a tip\` and *nie ma co [wychodzić]* \`it is not worth [leaving]\`, adjacent in the dependency tree. The nodes and arcs labelled with 1 are marked in **bold**.

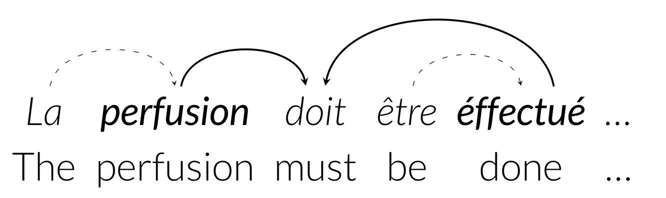

Figure 2: Extended encoding applied to a tree fragment with a disconnected French LVC.

Local model (basic encoding)

- **Input**: word vectors $\boldsymbol{w} = (\boldsymbol{w}_i \in \mathbb{R}^d)_{i=1}^n$, dependency graph $G = (V, E)$
- **Score**: given $(i, j) \in E$

$$\Phi(i,j) = \mathsf{MLP}^{(1)}([\boldsymbol{w}_i; \boldsymbol{w}_j]) \in \mathbb{R}^2 \tag{1}$$

- **Probability**:

$$P(\ell_E(i,j) \mid \boldsymbol{w}, G) = \mathsf{SoftMax}(\Phi(i,j)) \tag{2}$$

- **Prediction**: independently for each $(i, j) \in E$ based on P

Global model (extended encoding)

- **Compound labeling**: function $\ell\colon E \to \{1, \ldots, 8\}$ which encodes the labeling decisions $\ell_V(i)$, $\ell_E(i,j)$, and $\ell_V(j)$ for a given $(i, j) \in E$
  $\Longrightarrow$ allows to capture the relations between the adjacent labeling decisions
- **Node score.** Given $i \in V$:

$$\phi_V(i) = \mathsf{MLP}^{(2)}(\boldsymbol{w}_i)_1 \in \mathbb{R} \tag{3}$$

- **Compound score.** Given $(i, j) \in E$:

$$\phi_E(i,j) = \mathsf{TweakedMLP}^{(3)}([\boldsymbol{w}_i; \boldsymbol{w}_j]) \in \mathbb{R}^8 \tag{4}$$

	_,0,_	0,1,0	0,1,1	1,1,0	1,1,1
nie → ***ma***	0	-1	-1	-1	0
ma → ***co***	0	-1	-1	-1	1

	_,0,_	0,1,0	0,1,1	1,1,0	1,1,1
perfusion → *doit*	0	0	-1	1	-1
doit → ***effectué***	0	0	1	-1	-1

Table 1: Example scores which allow to capture (i) a 3-word and (ii) a disconnected VMWE

- **Global score.** Given a compound labeling ℓ:

$$\Phi(\ell) = \textstyle\sum_{i \in V} \phi_V(i)\ell_V(i) + \sum_{(i,j) \in E} \phi_E(i,j)_{\ell(i,j)} \tag{5}$$

- **Probability**:

$$P(\ell \mid \boldsymbol{w}, G) = \frac{\exp(\Phi(\ell))}{\sum_{\ell'} \exp(\Phi(\ell'))} \tag{6}$$

- **Prediction**: pick the global labeling which maximizes the global score
  $\Longrightarrow$ all the nodes on the VMWE border must be marked as its elements

System implementation

- **Input**: fastText [8] + hidden POS and dependency label embeddings
- **Training objective**: sum of the cross-entropies between the target and the estimated distributions for the individual arcs (`marginals` in the global model)
- **Frameworks**: `Keras` for the local model, `Haskell backprop` (automatic differentiation library) + `sgd` for the global model
- **Repository**: `https://github.com/kawu/vine`

Dataset

- German, French, and Polish datasets of PARSEME corpus, edition 1.1 [6]
- Tokenized, POS tagged, lemmatized, and enriched with dependencies

Pre-processing steps (automatic apart from the 3rd):

- Remove multiword tokens (e.g. the contraction *du* of *de le* \`of the\` in French)
- Add dummy root nodes (to enforce that dependency structures are trees)
- Add missing lemmas in French (for reliable comparison with ATILF [7])

Evaluation results

		DE			FR			PL			AVG		
		P	R	F	P	R	F	P	R	F	P	R	F
ATILF	MWE	71.56	46.71	**56.52**	82.69	71.38	76.62	85.23	68.35	**75.86**	79.82	62.15	**69.67**
	Token	76.43	45.72	57.21	85.73	72.96	78.83	88.69	67.9	**76.92**	83.61	62.19	**70.99**
Local	MWE	49.64	27.15	35.10	71.04	62.08	66.67	75.54	53.98	62.97	65.41	47.98	55.36
	Token	68.22	39.78	50.25	80.03	68.12	73.60	79.45	54.37	64.56	75.90	54.09	63.17
Global	MWE	68.48	47.70	56.24	84.92	70.75	**77.19**	80.83	64.66	71.84	78.08	61.04	68.52
	Token	72.74	47.83	**57.72**	86.84	73.24	**79.47**	83.13	66.19	73.69	80.90	62.42	70.47

Table 2: General results per language and system on DEV

		DE			FR			PL			AVG		
		P	R	F	P	R	F	P	R	F	P	R	F
ATILF	MWE	70.82	39.96	51.09	74.57	61.24	**67.25**	80.94	60.19	69.04	75.44	53.80	62.81
	Token	76.03	39.69	52.16	79.83	65.93	**72.22**	83.21	59.48	69.37	79.69	55.03	65.10
Local	MWE	54.36	26.31	35.45	60.26	55.42	57.74	74.46	60.00	66.45	63.03	47.24	54.00
	Token	70.3	36.82	48.38	73.96	62.08	67.50	78.95	59.57	67.90	74.48	52.82	61.81
Global	MWE	69.72	44.38	**54.23**	74.57	60.64	66.89	82.01	66.41	**73.39**	75.43	57.14	**65.02**
	Token	74.52	44.10	**55.41**	78.56	63.54	70.25	83.85	66.06	**73.90**	78.98	57.90	**66.82**

Table 3: General results per language and system on TEST

	Contin-uous	Discon-tinuous	Multi-token	Single-token	Seen-in-train	Unseen-in-train	Variant-of-train	Identical-to-train
ATILF	72.19	44.79	60.26	69.08	**82.15**	18.9	71.87	**92.72**
Local	56.68	47.96	56.37	0.0	72.29	29.59	68.06	75.88
Global	**72.58**	**53.30**	**62.67**	**69.89**	81.65	**32.28**	**74.07**	89.23

Table 4: MWE-based F-scores per VMWE challenge averaged over the three language test sets.

		VID	LVC.full	VPC.full	IRV	IAV
DE	ATILF	**39.29**	19.23	64.55	28.57	-
	Local	33.67	21.87	40.29	30.77	-
	Global	35.56	**22.95**	**72.40**	**32.84**	-
	#	37%	8%	42%	8%	0%
FR	ATILF	64.47	60.9	-	73.53	-
	Local	51.08	53.25	-	75.93	-
	Global	**66.12**	**61.29**	-	**78.47**	-
	#	43%	32%	0%	22%	0%
PL	ATILF	**46.73**	50.81	-	86.08	60.0
	Local	13.01	64.86	-	85.71	0.0
	Global	35.51	**65.62**	-	**87.32**	**69.57**
	#	14%	29%	0%	48%	6%

Table 5: MWE-based F-scores for the selected VMWE categories on the test sets.

	DE		FR	
	All	Dis.	All	Dis.
H-comb.	**60.71**	**57.53**	76.56	**67.23**
Global	56.24	51.47	**77.19**	64.84

Table 6: Comparison with H-combined [9] in terms of the MWE-based F-score (all and discontinuous VMWEs) on DEV.

	DE		FR	
	All	Dis.	All	Dis.
H-comb.	**59.29**	**55.00**	**70.97**	**63.90**
Global	58.05	47.49	68.59	58.15

Table 7: Comparison with H-combined on TEST (training on TRAIN+DEV).

- **Note**: for each language and VMWE category, 3 global models were trained and used to calculate ensemble node and compound scores
- **ELMo**: preliminary experiments on German show better perfomance on VIDs, worse on VPCs, and clear over-fitting

Conclusions & future work

- Dependency-based VMWE encoding method with high coverage
- (Close to) SOTA results despite a fairly simple and transparent neural architecture
- ◇ Obfuscate the architecture (contextualized word embeddings, BiLSTM, self-attention, higher-order factors, ...)
- ◇ Enhance the encoding schemata ($\Rightarrow$ support for overlapping VMWE occurrences, encoding VMWE categories)
- ◇ Extend the method to joint dependency parsing [10] and VMWE identification

References

[1] Bejcek, E. et al. (2012). Prague Dependency Treebank 2.5 -- a revisited version of PDT 2.0. **[2]** Abeillé, A. and Schabes, Y. (1989). Parsing Idioms in Lexicalized TAGs. **[3]** Nagy T., I. and Vincze, V. (2014). VPCTagger: Detecting Verb-Particle Constructions With Syntax-Based Methods. **[4]** Waszczuk, J. (2018). TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model. **[5]** Schneider, N. and Smith, N. A. (2015). A Corpus and Model Integrating Multiword Expressions and Supersenses. **[6]** Ramisch, C. et al. (2018). Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1). **[7]** Al Saied, H., Constant, M., and Candito, M. (2017). The ATILF-LLF System for Parseme Shared Task: a Transition-based Verbal Multiword Expression Tagger. **[8]** Mikolov, T. et al. (2018). Advances in Pre-Training Distributed Word Representations. **[9]** Rohanian, O. et al. (2019). Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. **[10]** Dozat, T. and Manning, C. D. (2017). Deep Biaffine Attention for Neural Dependency Parsing.