A Survey of MWE Identification Experiments: The Devil is in the Details

Carlos Ramisch, Abigail Walsh,
Thomas Blanchard, and Shiva Taslimipoor

MWE 2023 Workshop

Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues

Introduction

Full-text multiword expression identification – sequence annotation

→ Focus of recurrent shared tasks (ST): DiMSUM & PARSEME

Survey's goal

- Analyses MWE identification papers with experiments on data
- Look at methodological issues often seen as minor or omitted
- Hypothesis: these issues influence results and conclusions

Scope

Selection criteria:

- Available on the ACL Anthology
- Focus on MWE identification (Constant et al. 2017)
- Report experimental results
  - DiMSUM or PARSEME shared task or system description

    OR
  - report experiments on DiMSUM or PARSEME corpora

Scope

Selection criteria:

- Available on the ACL Anthology
- Focus on MWE identification (Constant et al. 2017)
- Report experimental results
    - DiMSUM or PARSEME shared task or system description

      OR
    - report experiments on DiMSUM or PARSEME corpora

Paper stats

- 40 papers

  → 4 overall ST papers

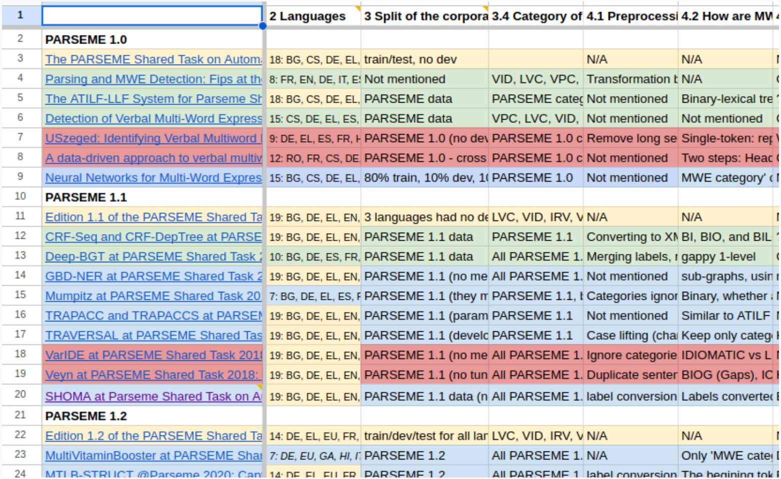
  → 27 ST system descriptions

  → 9 non-ST system descriptions

Questions

- Data
  - → Corpora
  - → Pre- and post-processing
  - → Sequence label encoding and decoding
- Evaluation
  - → Metrics
  - → Significance of comparisons
  - → Error analysis

Working table



1		2 Languages	3 Split of the corpora	3.4 Category of	4.1 Preprocessi	4.2 How are MW
2	**PARSEME 1.0**					
3	The PARSEME Shared Task on Autom	18: BG, CS, DE, EL,	train/test, no dev		N/A	N/A
4	Parsing and MWE Detection: Fips at th	8: FR, EN, DE, IT, E	Not mentioned	VID, LVC, VPC,	Transformation t	N/A
5	The ATILF-LLF System for Parseme Sh	18: BG, CS, DE, EL,	PARSEME data	PARSEME categ	Not mentioned	Binary-lexical tre
6	Detection of Verbal Multi-Word Express	15: CS, DE, EL, ES,	PARSEME data	VPC, LVC, VID,	Not mentioned	Not mentioned
7	USzeged: Identifying Verbal Multiword	9: DE, EL, ES, FR, H	PARSEME 1.0 (no dev	PARSEME 1.0 c	Remove long se	Single-token: rep
8	A data-driven approach to verbal multiv	12: RO, FR, CS, DE,	PARSEME 1.0 - cross	PARSEME 1.0 c	Not mentioned	Two steps: Head
9	Neural Networks for Multi-Word Expres	15: BG, CS, DE, EL,	80% train, 10% dev, 1(	PARSEME 1.0	Not mentioned	MWE category' c
10	**PARSEME 1.1**					
11	Edition 1.1 of the PARSEME Shared Ta	19: BG, DE, EL, EN,	3 languages had no de	LVC, VID, IRV, V	N/A	N/A
12	CRF-Seq and CRF-DepTree at PARSE	19: BG, DE, EL, EN,	PARSEME 1.1 data	PARSEME 1.1	Converting to XM	BI, BIO, and BIL
13	Deep-BGT at PARSEME Shared Task 2	10: BG, DE, ES, FR,	PARSEME 1.1 data	All PARSEME 1.	Merging labels, r	gappy 1-level
14	GBD-NER at PARSEME Shared Task 2	19: BG, DE, EL, EN,	PARSEME 1.1 (no me	All PARSEME 1.	Not mentioned	sub-graphs, usin
15	Mumpitz at PARSEME Shared Task 20	7: BG, DE, EL, ES, F	PARSEME 1.1 (they m	PARSEME 1.1, t	Categories ignor	Binary, whether a
16	TRAPACC and TRAPACCS at PARSEM	19: BG, DE, EL, EN,	PARSEME 1.1 (param	PARSEME 1.1	Not mentioned	Similar to ATILF
17	TRAVERSAL at PARSEME Shared Tas	19: BG, DE, EL, EN,	PARSEME 1.1 (develo	PARSEME 1.1	Case lifting (cha	Keep only categ
18	VarIDE at PARSEME Shared Task 201	19: BG, DE, EL, EN,	PARSEME 1.1 (no me	All PARSEME 1.	Ignore categorie	IDIOMATIC vs L
19	Veyn at PARSEME Shared Task 2018:	19: BG, DE, EL, EN,	PARSEME 1.1 (no tun	All PARSEME 1.	Duplicate senter	BIOG (Gaps), IC
20	SHOMA at Parseme Shared Task on A	19: BG, DE, EL, EN,	PARSEME 1.1 data (n	All PARSEME 1.	label conversion	Labels convertec
21	**PARSEME 1.2**					
22	Edition 1.2 of the PARSEME Shared Ta	14: DE, EL, EU, FR,	train/dev/test for all lar	LVC, VID, IRV, V	N/A	N/A
23	MultiVitaminBooster at PARSEME Shar	*7: DE, EU, GA, HI, I*	PARSEME 1.2	All PARSEME 1.	N/A	Only 'MWE cate
24	MTLB-STRUCT @Parseme 2020: Can	14: DE, EL, EU, FR,	PARSEME 1.2	All PARSEME 1.	label conversion	The begining tok

Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues

Corpus splits

Shared tasks

- DiMSUM: 3 domains, 1 lang, train + test
- PARSEME 1.0: news, 18 lang, train + test
- PARSEME 1.1: news, 19 lang, train + test + dev (16 lang)
- PARSEME 1.2: news, 14 lang, train + test + dev

  → Biased split: focus on unseen MWEs

Corpus use

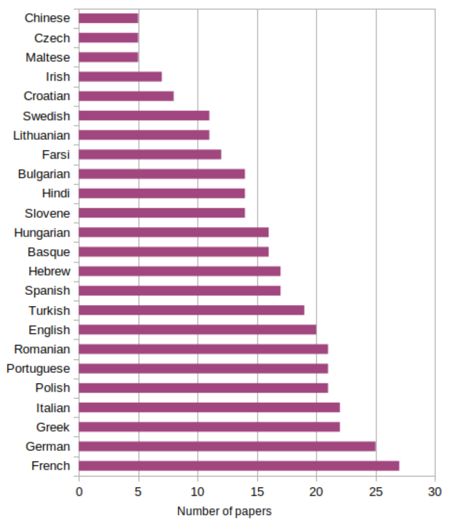
- Training corpus unused: 4/36 papers
  - → External resources (2 papers), other corpora (2 papers)
- Development corpus not provided:
  - → Custom train-dev set: 6/36 papers
  - → Cross-validation: 3/36 papers
  - → Dev on another language: 2/36 papers
  - → Dev corpus not mentioned: 3/36 papers

Corpus use

- Training corpus unused: 4/36 papers
  - → External resources (2 papers), other corpora (2 papers)
- Development corpus not provided:
  - → Custom train-dev set: 6/36 papers
  - → Cross-validation: 3/36 papers
  - → Dev on another language: 2/36 papers
  - → Dev corpus not mentioned: 3/36 papers

Recommendation
Always mention development data

Languages


Chinese
Czech
Maltese
Irish
Croatian
Swedish
Lithuanian
Farsi
Bulgarian
Hindi
Slovene
Hungarian
Basque
Hebrew
Spanish
Turkish
English
Romanian
Portuguese
Polish
Italian
Greek
German
French
0
5
10
15
20
25
30
Number of papers

Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues

Pre-processing

Variants of BIO-style encoding: 12/36 papers

DiMSUM	The	staff	leaves	a	lot	to	be	desired	.
	O	O	B	b	i_	I_	I_	I_	O
PARSEME	I	did	a	lot	of	study	and	research	.
	*	1:LVC;2:LVC	*	*	*	1	*	2	*

- Gaps: 12/36 papers account for gaps
- Nesting and overlaps

→ Ignored, handled by modifying BIO-style

→ Kept the tags as they are, dependency graphs

→ No mention (most papers)

Post-processing

Conversion from BIO-style

- Combination heuristics (7/36 papers)
  - → B-labelled and I-labelled words matched
  - → Standalone I-labelled ignored
- Greedy-matching algorithm (1/36 paper)
- Viterbi decoding (1/36 paper)
- Conditional random fields (8/36 papers)
- Dependency trees (2/36 papers)
  - → Elements of MWE assumed to be nodes in the same subtree

Post-processing

Conversion from BIO-style

- Combination heuristics (7/36 papers)
  - → B-labelled and I-labelled words matched
  - → Standalone I-labelled ignored
- Greedy-matching algorithm (1/36 paper)
- Viterbi decoding (1/36 paper)
- Conditional random fields (8/36 papers)
- Dependency trees (2/36 papers)
  - → Elements of MWE assumed to be nodes in the same subtree

Recommendation
Explicitly report all pre- and post-processing + MWE encoding

Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues

Evaluation metrics

DIMSUM exact match and linked-based P, R and F1

PARSEME MWE-based and token-based P, R and F1

PARSEME focused measures:

- Seen/Unseen: focus of 9 papers
- Diversity: 2 PARSEME papers
- Discontinuity: focus of 5 papers

Evaluation metrics

DIMSUM exact match and linked-based P, R and F1

PARSEME MWE-based and token-based P, R and F1

PARSEME focused measures:

- Seen/Unseen: focus of 9 papers
- Diversity: 2 PARSEME papers
- Discontinuity: focus of 5 papers

Recommendation
Focused measures help highlight system strengths and limitations

Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues

Compare systems A and B

- Test set
  - $x = x^{(1)} \dots x^{(m)}$ – m input sentences
  - $y = y^{(1)} \dots y^{(m)}$ – m reference MWE annotations
- Method :
  1. Apply A to x to obtain $\hat{y}_A$, compare to y
  2. Calculate evaluation metric $M(A, x, y)$ (e.g. MWE-based F1)
  3. Do the same for B, obtain $M(B, x, y)$
  4. Calculate difference (effect)

$$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)$$

- $\delta_{A-B}(x, y) > 0 \implies A$ better than B?

Hypothesis testing

- $H_0 : \delta(X, Y) \leq 0 \implies$ if true, then A not better than B
- $H_1 : \delta(X, Y) > 0$
- $X, Y \rightarrow$ random variables, all possible test sets
  - Of which x, y is an m-sized sample
- Reject $H_0 \implies$ significant difference between the systems
- **P-value**: probability of observing $\delta_{A-B}(x, y)$ while H_0 is true:
  - $p - value = P[\delta(X, Y) \geq \delta_{A-B}(x, y) | H_0]$
  - probability to reject H_0 when it is true

Bootstrap p-value (Berg-Kirkpatrick et al. 2012)

Input

- Test set $x = x^{(1)} \dots x^{(m)}, y = y^{(1)} \dots y^{(m)}$,
- Predictions $\hat{y}_A^{(i)}$ and $\hat{y}_B^{(i)}$ of systems A and B
- Evaluation metric $M(\cdot)$

```
deltaobs = M(A,x,y) - M(B,x,y)  # observed difference
for i in range(R) :             # R constant 10k
  xsample, ysample = sample(x,y,m)  # m with repetition
  deltasample = M(A,xsample,ysample) - M(B,xsample,ysample)
  if deltasample > 2 * deltaobs :
        r = r + 1
pvalue = r/R                    # % of surprising results
return pvalue
```

Significance analysis

- Only 2/40 papers report significance
- Our tool estimates p-values for two CUPT predictions
  - → `https://gitlab.com/parseme/significance`
- We compare all system pairs and metrics of PARSEME 1.2
  - → 2,728 p-values in total
  - → 783 above the $\alpha = 0.05$ threshold (29%)

P-values for MWE-based F1 in Swedish

Systems		TRAVIS-multi	Seen2Unseen	TRAVIS-mono
	F1	**0.6911**	**0.6892**	**0.6709**
MTLB-STRUCT	**0.7158**	0.025	0.038	0.0
TRAVIS-multi	**0.6911**		0.464	0.081
Seen2Unseen	**0.6892**			0.103

P-values for MWE-based F1 in Swedish

Systems		TRAVIS-multi	Seen2Unseen	TRAVIS-mono
	F1	**0.6911**	**0.6892**	**0.6709**
MTLB-STRUCT	**0.7158**	0.025	0.038	0.0
TRAVIS-multi	**0.6911**		0.464	0.081
Seen2Unseen	**0.6892**			0.103

Recommendation
Systematically calculate/report p-values for model comparison

Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues

Error Analysis

- 33/36 papers report some error analysis
- 11/36 report MWE category or cross-language analyses
- Heterogeneous analyses
    - → Discontinuities, seen/unseen
    - → POS sequences, syntactic structure
    - → Ablation, role of external lexicons
    - → Pre-trained embeddings, tagging schemes

Error Analysis

- 33/36 papers report some error analysis
- 11/36 report MWE category or cross-language analyses
- Heterogeneous analyses
  - → Discontinuities, seen/unseen
  - → POS sequences, syntactic structure
  - → Ablation, role of external lexicons
  - → Pre-trained embeddings, tagging schemes

Recommendation
Error analyses uncover interesting phenomena for future work

Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues

Recommendations

We advocate reporting on experimental choices:

- corpus constitutions and selections
- pre- and post-processing
- evaluation metrics and significance testing of performance
- error analysis

We encourage focused measures that facilitate error analysis

We propose a tool to predict p-values from 2 CUPT predictions

Open issues

- Hyper-parameter tuning
  - → Selection of the data
  - → Strategy (e.g. grid search, random, etc.)
- Should manual evaluation of detected MWEs be performed?
- New evaluation protocols
  - → e.g. are some MWE categories more important than others?

Thanks! Questions?

Acknowledgments

- SELEXINI project (ANR-21-CE23-0033-01)
- Ulysses PHC grant (45358QD)
- Marius Thorre, Silvio Cordeiro, Agata Savary, reviewers