EACL'23 – The 17th Conference of the European Chapter of the
Association for Computational Linguistics – May 2-6, 2023, Croatia

Predicting Compositionality of Verbal Multiword Expressions in Persian

Mahtab Sarlak, Yaldasadat Yarandi & Mehrnoush Shamsfard

NLP Research Laboratory, Shahid Beheshti University, Iran
ma.sarlak@mail.sbu.ac.ir, y.yarandi@mail.sbu.ac.ir, m-shams@sbu.ac.ir

Summery

Introduction: background information, goals

Related work

VMWE Identification

Predicting the Compositionality of VMWEs

Results and discussion

Conclusion: conclusion, limitations

Introduction

- The violation of the compositionality
- Discontiguous VMWEs
- The assignment of grammatical roles to certain word sequences can be entirely dependent on the sense of the words and the context in which they are used.

Introduction (cont.)

- VMWEs can possess both idiomatic and literal meanings, leading to syntactic ambiguity.
- Problematic for embedding vectors that accurately capture semantic meaning.
- One of the defining challenges of VMWEs is their compositional nature.

Introduction (cont.)

- Identifying VMWEs in Persian sentences as a sequence labeling problem
  - Non-Contextual
  - Contextual
- Creating word embeddings that better capture the semantic properties
  - Analyzing the semantic similarity between its components and the expression

Related work

- Identification of verbs in Persian language sentences (Chaghari and Shamsfard, 2013)
- Recognizing MultiWord Expressions in Persian (Saljoughi, 2016)
- Automatic identification of Persian light verb constructions (Salehi, 2012)
- Shoma at parseme shared task on automatic identification of vmwes: Neural multiword expression tagging with high generalisation (Taslimipoor and Rohanian, 2018)

Related work (cont.)

- Salehi et al. (2015) compared word2vec and MSSG models to predict compositionality of MWEs in English and German datasets.
- Combining string similarity with word embeddings was found to be comparable to existing methods (Salehi and Cook, 2013).
- Nandakumar et al. (2018) used different types of embeddings and found that word2vec performed the best for English MWEs.
- Cordeiro et al. (2019) proposed preprocessing MWEs into a single unit for model training, which requires a comprehensive list of MWEs.

VMWE Identification

Train Datasets:

- Parseme corpus train,development-set(Savary et al., 2017)
  - 3226 sentences
- Persian Dependency Treebank (Rasooli et al., 2013)
  - 30,000 sentences
  - rule-based strategy
    - dependency tree tags
    - Resulting in 32056 VMWEs in the training set of the corpus
  - 1000 sentences tagged manually

Test Dataset:

- Parseme corpus test-set

VMWE Identification (cont.)

non-contextual method

- Created a dataset of VMWEs by collecting all compound verbs in FarsNet (Shamsfard, 2007)
  - Extracted 21,462 VMWEs
- Extracted n-grams (for n=2, 3, 4) in a sentence and searched for the presence of all components of a multi-word verb within the n-gram.
  - not all cases that are found are VMWEs, and not all VMWEs can be found

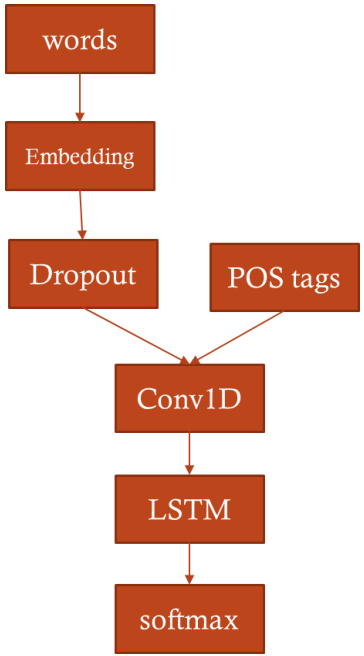
VMWE Identification(cont.)

contextual method

- IOB-like labelling format
- Sentences containing two VMWEs with mixed components were removed for simplicity

VMWE Identification (cont.)

- Using the combination of a ConvNet and a long short-term memory network
- Using a a pre-trained neural model on Persian Texts (ParsBERT)



Compositionality prediction

- Determine degree of compositionality of a multiword expression through semantic similarity between its components and the expression
- Investigate six metrics for determining compositionality of VMWEs

Compositionality prediction

- Four corpora for training word2vec and fasttext embedding models for VMWEs.
- Metrics focus on calculating the similarity between the VMWE's components and the VMWE itself, followed by determining a threshold.
- Cosine similarity is used for all similarity calculations.
- Embedding models are also trained on the original corpora to obtain the embedding vectors of all VMWE components.

Compositionality prediction

- Syn_Sim
- If the sim_syn_vmwe > sim_syn_combined => the constructed VMWE's vector provides a better representation

$$combined_{vector} = \sum_{i=1}^{N} w_i \quad (1)$$

$$sim\_syn\_vmwe = cos(vmwe, syn\_verb_1) \quad (2)$$

$$sim\_syn\_combined = \cos(combined_{vector}, syn\_verb_1) \quad (3)$$

Compositionality prediction

- **Direct_pre**
- Compositional VMWEs tend to have a similar context with their components
- Calculate the similarity between the VMWE's embedding vector and the 'combined' vector of its components

Compositionality prediction

- **Direct_post**
- The similarity between the vector embedding of a VMWE and each of its components.

$$direct\_post = \alpha \cos(vmwe, w1) + (1-\alpha)\cos(vmwe, w2)$$

Compositionality prediction

- **DFsum**
- The similarity between the vector embedding of a VMWE and the element-wise sum of normalized vectors of its components is computed

$$combined\_vector\_norm = \sum_{i=1}^{N} \frac{w_i}{|w_i|} \quad (6)$$

$$DFsum = cos(vmwe, combined\_vector\_norm) \quad (7)$$

Compositionality prediction

- **DFcomp**
- The similarity between the VMWE's components' word vectors is computed.
- **DFsing**
- The similarity between the vector embedding of a VMWE and the vector of the most similar single word.

Compositionality prediction

Dataset :

- Bijankhan, HmBlogs, PARSEME, and PerDT
  - We used the first 1 million sentences of Hmblogs.
- Compositional and non-compositional VMWE dataset
  - The dataset was gathered through the works of Karimi (1997) and Sharif (2017).
  - Consists of 33 compositional and 22 non-compositional annotated verbs in infinitive form.

Results and discussion

	Token-based			VMWE-based			Sentence-based
	precision	**recall**	**f1 score**	**precision**	**recall**	**f1 score**	**accuracy**
Non-Contextual	-	-	-	34.19%	43.71%	38.36%	-
LSTM	61.50% 69.95% 63.39%	49.23% 50.40% 51.03%	54.71% **58.59%** 56.54%	72.00% 85.52% 72.34%	60.07% 63.67% 61.07%	65.50% **72.99%** 66.23%	51.11% **58.05%** 53.61%
BERT	94.04% 90.34% 94.88%	84.25% 74.54% 77.86%	**88.87%** 81.68% 85.53%	92.37% 91.43% 93.25%	85.99% 77.90% 79.09%	**89.07%** 84.13% 85.59%	**71.38%** 63.88% 68.61%

Results and discussion

seen and unseen verbs

- seen verbs as verbs whose exact forms (like their persons, tenses etc.) exist in the train set.
- turn the core (the main verb) of all verbal expressions in the test and train set to their infinitive form and then check whether the expression exists in the train set.

	Seen proportion	**Correct detection of the seen verbs**	**Correct detection of the unseen verbs**
Method 1	33.33%	89.00%	62.56%
Method 2	73.12%	80.42%	46.75%

Results and discussion

- For analyzing the compositionality of VMWE, only the word2vec model trained on Hmblog, the largest corpus, is considered.

Criterion	threshold	accuracy
Direct pre	0.23	**0.709**
Direct post	0.27	0.655
DFcomp	0.23	0.618
DFsum	0.23	**0.709**

Results and discussion

VMWE	syn_verb	sim_syn_vmwe	sim_syn_combined
در_نظر_گرفتند (in consider got => considered)	شمردن (considering)	0.81	0.62
خشمگین_شده (angry become => get angry)	برافروختن (getting angry)	0.88	0.63
بیان_می_کرد (expression was doing => was expressing)	فرمودن (saying)	0.83	0.50

Results and discussion

non-compositional	Direct_pre	DFcomp	freq	compositional	DFcomp	Direct_pre	freq
چشم_زدن (eye hitting => jinxing)	**0.23**	0.22	7	نگاه_کنید (look do => look)	0.30	0.37	296
فریب_خورده (deception ate => deceived)	**0.25**	0.40	28	تغییر_کند (change do => change)	0.33	0.43	130
دوست_دارم (friend have => to like)	**0.10**	0.56	1032	خاک_کرد (soil did => buried)	0.16	0.23	3
شکست_خورده (failure ate => failed)	0.17	0.51	132	فکر_کنید (think do => think)	0.24	0.40	258
زمین_خوردن (land eating => falling down)	0.13	0.29	50	قرار_دادن (put have => putting up)	0.32	0.38	1806
چانه_زدن (chin hiting => to bargaining)	0.14	0.4	62	آمده_به_دنیا (to world came => born)	0.25	0.51	105

Conclusion

- Predicting the compositional nature of VMWEs in Persian
- The method involved automatic identification of VMWEs, creating word embeddings to capture their semantic properties, and using multiple criteria to determine their degree of compositionality.
- A fine-tuned BERT model outperformed the BiLSTM model with an F1 score of 89%
- Resulting in an accuracy of 70.9% on a collected dataset of expert-annotated compositional and non-compositional VMWEs
- Limited annotated dataset
- High prevalence of VMWEs in Persian
- low-resource language

Thank You

In memory of Maloos