Using Noun Similarity to Adapt an Acceptability Measure for Persian Light Verb Constructions

Shiva Taslimipoor[*], **Afsaneh Fazly**[†], **Ali Hamzeh**[*]

[*]Electrical and Computer Engineering
Shiraz University
Shiraz, Iran
sh.taslimi@gmail.com, ali@cse.shirazu.ac.ir

[†]Institute for Research in Fundamental Sciences (IPM)
Tehran, Iran
afsaneh.fazly@gmail.com

Abstract

Light verb constructions (LVCs), such as *take a walk* and *make a decision*, are a common subclass of multiword expressions (MWEs), whose distinct syntactic and semantic properties call for a special treatment within a computational system. In particular, LVCs are formed semi-productively: often a semantically-general verb (such as *take*) combines with a number of semantically-similar nouns to form semantically-related LVCs, as in *make a decision*/*choice*/*commitment*. Nonetheless, there are restrictions as to which verbs combine with which class of nouns. A proper computational account of LVCs is even more important for languages such as Persian, in which most verbs are of the form of LVCs. Recently, there has been some work on the automatic identification of MWEs (including LVCs) in resource-rich languages, such as English and Dutch. We adapt such existing techniques for the automatic identification of LVCs in Persian, an under-resourced language. Specifically, we extend an existing statistical measure of the acceptability of English LVCs (Fazly et al., 2007) to make explicit use of semantic classes of noun, and show that such classes are in particular useful for determining the LVC acceptability of new combinations.

Keywords: Multiword Expressions, Semi-productivity, Persian Light Verb Constructions

1. Introduction

A Multiword Expression (MWE) consists of two or more words that together have a meaning different from the composition of the component meanings. Light Verb Constructions (LVCs) are a subtype of verbal MWEs, formed from the combination of a semantically-general *basic* verb with a content-bearing word. Basic verbs are high-frequency highly-polysemous verbs that express events or actions central to human experience, e.g., *take* in English, and *zadan* (lit. 'to hit') in Persian.[1] LVCs, like other types of MWEs, require special treatment within a computational system, such as machine translation, summarization, and parsing. For example, an automatic parser should realize that in *take a walk*, *walk* is not a direct object of *take*, and that *take a walk* is a complex predicate. Despite their idiosyncratic behavior, LVCs tend to be semi-productive, in that semantically-similar words tend to combine with the same verb to form LVCs, as in *take a walk/hike/stroll* in English, and *târ/setâr/âhang zadan* ('to play music/a musical instrument') in Persian.

LVCs are very common and highly productive in Persian: Most verbs in Persian are of the complex form, and they greatly outnumber the single-word verbs of this language (Khanlari, 1973). Nonetheless, there are restrictions on what kind of nouns a verb can combine with to form acceptable LVCs. For example, although the verb *zadan* occurs with a wide range of nouns, it tends to productively combine with certain semantic classes of nouns. Table 1 provides some examples, most of which are taken from Mansoory and Bijankhan (2008). Persian LVCs have received a lot of attention in the linguistics literature (Karimi, 1997; Dabir-Moghaddam, 1997; Megerdoomian, 2004). However, there has not been much computational work on the automatic treatment of these expressions; though see Mansoory and Bijankhan (2008) and Rouhizadeh et al. (2010), for very preliminary studies. In particular, the Persian language lacks large-scale lexical resources which are necessary for the development of scalable Natural Language Processing (NLP) systems. The automatic identification of LVCs is an important first step in the creation of such resources for Persian.

Much recent research has looked into the extraction of MWEs (Baldwin and Villavicencio, 2002), as well as learning about their semantics (McCarthy et al., 2003; Bannard et al., 2003; Baldwin et al., 2003; Fazly et al., 2009; Fazly and Stevenson, 2007). However, only a few studies have focused on the semi-productivity of MWEs, including English LVCs as in Stevenson et al. (2004) ,Fazly et al. (2007), and English verb-particle constructions such as *finish up*, as in Villavicencio (2003). In particular, Fazly et al. (2007) propose a probabilistic measure for determining the acceptability of a combination of a verb and a noun as an LVC. This measure shows reasonably good correlations/agreements with human judgments, both in determining the degree of acceptability of an individual verb+noun, and in predicting the level of productivity of verb plus a semantic class of nouns.

We extend the probabilistic measure of Fazly et al. (2007) in a few directions: First, we examine the generalizability of the measure by testing it on Persian LVCs. Second,

[1] We follow Fazly (2007) and refer to these verbs as *basic*.

Table 1: The basic verbs,*zadan*, and LVC examples.

Semantically-similar nouns that form LVCs with *zadan* (lit. ‘to hit’)		Meaning of *zadan*
piâno (piano) / *târ zadan* (fiddle) / *âhang* (music)	+ *zadan*	‘to play’
telegrâf (telegraph) / *email* (email) / *fâx* (fax)	+ *zadan*	‘to send’
harf (talk) / *faryâd* (shout) / *soot* (whistle)	+ *zadan*	‘to do’
rang (paint) / *roghan* (oil) / *âb* (water)	+ *zadan*	‘to cover with’
sadameh (hurt) / *lagad* (kick) / *zarbeh* (stroke)	+ *zadan*	‘to hit’

we extend the evaluation of the measure with regard to its success at predicting productivity. Specifically, we replace the direct corpus-based estimation of the statistical components of the measure by similarity-based estimations, and then use these new estimations to predict the LVC acceptability of low-frequency expressions. We believe this is in particular useful for determining productivity. That is, we expect that knowledge about the productivity of a given verb in combining with members of a semantic class should help predict the acceptability of a novel (or a very low-frequency) combination.

2. LVC Acceptability Measure

2.1. The Base Measure

Fazly et al. (2007) measure the acceptability of a combination of a verb V and a noun N as an LVC — which we call LVC(N, V) — by estimating the joint probability $Pr(N, LV, LVC)$ as in:

$$\text{LVC}(N, V) = Pr(N, LV, LVC) = \\ Pr(N)\ Pr(LVC|N)\ Pr(LV|N, LVC) \qquad (1)$$

The first factor $Pr(N)$ is estimated by the relative frequency of occurrence of the noun N in a corpus. The second factor $Pr(LVC|N)$ is the probability with which N forms an LVC with any verb, and is estimated as the relative frequency of N appearing in the prototypical LVC pattern (“V a/an N” in English) across a few known basic verbs. The third component $Pr(LV|N, LVC)$ is the probability that the noun N forms an LVC with the given verb LV, and is estimated similarly to the second factor, but only looking at the verb LV. We use this as our base measure to detect the acceptability of LVCs in Persian. We consider the prototypical Persian LVC pattern to be “N V”, where N immediately precedes V (note that Persian is a verb-final language).[2]

2.2. A Similarity-based Estimation

The success of LVC will largely depend on the reliability of the frequency estimates. The measure thus may not work as well on low-frequency items. We draw on the semi-productivity of LVCs, and use semantic similarity among nouns to provide more reliable estimates of the components of the measure for low-frequency items. We assume that, if a given noun N in a low-frequency (or novel) candidate N+V is semantically similar to nouns that tend to form (high-frequency) LVCs with the verb V, then it is more likey that N+V is an acceptable LVC. We propose a similarity-based measure, LVC$_{\text{SIM}}$, that estimates the above joint probability for a low-frequency target candidate N+V as follows: (i) find a set of k nouns appearing in high-frequency candidates that are semantically similar to N; (ii) estimate each component of joint Pr for the target by taking the average of the estimates of that component for the k similar nouns. Next, we explain how we measure semantic similarity among nouns.

2.3. Measuring Semantic Similarity of Nouns

We use a distributional vector-space method for measuring the semantic similarity among nouns, using the Gensim package for extracting the distributional vectors (Řehůřek and Sojka, 2010). We experiment with two kinds of dimension words: (i) the 1000 most frequent nouns (after removing the 100 highest-frequency nouns and some proper names as non-informative) referred to as the noun vectors; and (ii) the 100 most frequent verbs, expecting to see many of the basic verbs in this group, referred to as the verb vectors.

To construct the noun vectors, we take the context of each target word to be the dimension nouns within a window size of 10 around the target, a common window size used in many previous studies. The use of verb vectors is inspired by the semi-productivity patterns of LVC formation: We expect semantically-related nouns to have similar patterns of association to the high-frequency dimension verbs. For example, the semantically-related nouns in examples (1) and (2) below show consistent patterns in terms of whether they form acceptable combinations with the three high-frequency verbs of *dâdan* (lit. ‘to give’), *yâftan* (lit. ‘to find’), and *khordan* (lit. ‘to eat/collide’).

1. (a) *afzâyesh/kâhesh/taghlil dâdan*[3]
   (b) *afzâyesh/kâhesh/taghlil yâftan*
   (c) *??afzâyesh/kâhesh/taghlil khordan*
2. (a) *ghasam/sogand dâdan*[4]
   (b) *??ghasam/sogand yâftan*
   (c) *ghasam/sogand khordan*

To construct verb vectors, we consider the context of a target word to be the immediately following word. In doing so, we assume that when this context word for a noun is one of the high-frequency dimension verbs, it helps find semantically-related nouns that are also similar in terms of

[2] Note that it is possible for the two components of a Persian LVC to be separated (Karimi-Doostan, 2011). We thus expect our extraction method to identify some (but not all) LVC instances.

[3] *afzâyesh* means ‘growth’, *kâhesh* and *taghlil* mean ‘decrease/decline’.

[4] *ghasam* and *sogand* mean ‘vow/pledge’.

which basic verbs they are more likely to combine with.[5] We calculate each dimension value using positive pointwise mutual information (PPMI) to measure the association strength between the target and the dimension word. PPMI is calculated by replacing the negative PMI values by zero, and is shown to be an effective weighting technique (Bullinaria and Levy, 2007). We use cosine to compute the similarity of each pair of vectors.

3. Experimental Setup

3.1. Corpus and Experimental Expressions

There are a limited number of basic verbs that form LVCs in Persian. For our experiments here, we choose five such verbs that are common and accompany a wide range of nouns in LVCs, namely, *zadan* (‘to hit’), *khordan* (‘to eat/collide’), *gereftan* (‘to take’), *dâdan* (‘to give’), and *gozâshtan* (‘to put’).

We extract our candidate expressions from Bijankhan,[6] a small corpus of about 2.6 million manually part-of-speech tagged words of Persian text. We extract all occurrences of a basic verb and its preceding noun. This simple extraction technique is very noisy, and results in many meaningless and erroneous expressions. We filter out noise by stop word elimination, stemming, and also excluding extremely low frequency expressions — that we take to be those with frequency < 5 in a large corpus, Hamshahri (explained below). These candidates are annotated by three Persian native speakers as being LVC or non-LVC. We use the majority label assigned to each candidate as its ‘true’ label (for evaluation). The final list of candidates contains 1098 expressions, including 547 LVCs.

Bijankhan is a small corpus. We thus use a larger corpus, Hamshahri,[7] for a more reliable estimation of the LVC measures, and for constructing reliable distributional vectors for measuring noun similarity. Hamshahri contains about 110 million words of untagged newswire text.

3.2. Evaluation

We divide our candidates into two groups according to their frequencies: Low-frequency (LF) items are those with frequency lower than 10, and everything else is considered high frequency (HF). We also randomly divide the expressions into DEVelopment and TEST portions, such that each portion includes more or less the same number of HF and LF LVCs: TEST contains 478 HF expressions (256 LVCs), and 70 LF expressions (17 LVCs); DEV is similar. We use DEV to find the best values for the k parameter.

We report our results on TEST (results on DEV have similar trends). We first compute the base LVC measure for all the 478 HF candidates in TEST. We then find k (set to 3 and 5) similar nouns appearing in an HF candidate for each of the nouns in the 70 LF candidates, and calculate LVC_{SIM} for these low-frequency items. We use two standard evaluation measures: the 11-point interpolated average precision (11-pt IAP), and the precision–recall curve. For these, we rank the candidates according to the score that a measure assigns to them, and calculate the interpolated precision at the 11 recall values of 0, 10%, $\cdots$, 100%. We compare the performance of LVC_{SIM} with that of LVC on LF items. We also compare the performances with that of a random baseline, Rand. Rand is the average interpolated precision at each recall across 100 randomly generated ranked lists of the candidate expressions under evaluation.

Table 2: 11-pt IAP for ALL and LF expressions in TEST. ‘DIM: Noun’ (‘DIM: Verb’) means dimensions of vector space models are nouns (verbs).

			LVC_{SIM}			
		Base	DIM: Noun		DIM: Verb	
	Rand	LVC	k=3	k=5	k=3	k=5
LF	30.5	53.9	**63.9**	50.8	**69**	**69.3**
ALL	51.8	80.3				

4. Results

Table 2 shows the 11-pt IAP for LF and ALL TEST expressions. Results on LF TEST expressions are given for both types of dimensions (nouns and verbs), and for the two values of k (3 and 5). First, we compare the performance of LVC to that of Rand: LVC shows a notable improvement over Rand, both on LF and on ALL expressions. These results suggest that this measure, originally proposed by (Fazly et al., 2007) for English LVCs, is sufficiently language independent, and can easily be extended to a different language, such as Persian. [8] The only thing we had to change to make the measure applicable to Persian LVCs was to come up with the prototypical pattern for LVCs in Persian.

Next, we look at the performance of LVC_{SIM} on LF expressions; since LVC_{SIM} is adapted to improve identifying LF LVCs. Improvements of LVC_{SIM} over LVC are shown in bold. Performance of LVC_{SIM} shows a great improvement over LVC on LF expressions in 3 out of 4 cases. Interestingly, we get much better results when we use verbs as dimensions, reinforcing our original motivation that a verb-based vector space might better capture similarity relevant to LVC formation. Figure 1 depicts the precision–recall curves on LF, for Rand, LVC, and for similarity-based measures with nouns and verbs as dimensions ($k = 3$). LVC_{SIM} with verb dimensions performs best, placing most LVCs at the top of the ranked list. Even though LVC_{SIM} with noun dimensions has a higher IAP compared to LVC, the curves show that the former performs better only at higher levels of recall (further down the ranked list), and that the base measure places many more LVCs at the top of the list.

5. Conclusions

In this study, we have examined the applicability of an existing measure of LVC acceptability, originally proposed for English LVCs (Fazly et al., 2007), to be used for identifying LVCs in a different language, here Persian. Drawing

[5] Although the verb component of an LVC is sometimes argued to be semantically empty, in many cases the verbs contributes some aspects of meaning to the expression (Karimi, 1997).

[6] http://ece.ut.ac.ir/dbrg/bijankhan/

[7] http://ece.ut.ac.ir/dbrg/hamshahri/

[8] Though Persian and English are genetically-related: Persian is Indo-Iranian, a branch of the Indo-European language family.

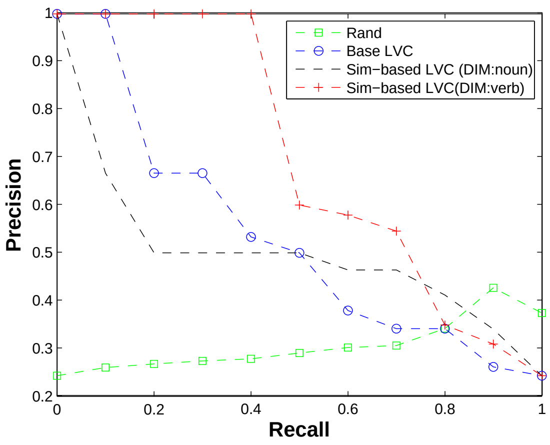

Figure 1: A comparison of the precision–recall curves.

on the semi-productivity of LVCs, we have also proposed a new similarity-based way of estimating the components of this measure.

Our results show that the original measure indeed works well on Persian LVCs (with an IAP of 80.3 compared to that of 51.8 for the baseline). Interestingly, our similarity-based measures outperform the original measure in determining the LVC acceptability of low-frequency expressions: the similarity-based measures result in 10% to 15% absolute increase in the IAP over the original LVC measure. These findings suggest that drawing on semantic classes of nouns is in particular helpful in the identification of low-frequency (or novel) LVCs.

Our ongoing work focuses on extending our similarity-based measures to better estimate the LVC acceptability of expressions of all frequency ranges. In addition, we are currently annotating more candidate expressions, in order to evaluate our measures on larger data sets that also include many more low-frequency LVCs.

6. References

T. Baldwin and A. Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002), Taipei, Taiwan*.

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An emprical model of multiword expression decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96.

C. Bannard, T. Baldwin, and A. Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72.

J.A. Bullinaria and J.P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*.

M. Dabir-Moghaddam. 1997. Compound verbs in Persian. *Studies in the Linguistic Sciences*, 27(2):25–59.

A. Fazly and S. Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.

A. Fazly, S. Stevenson, and R. North. 2007. Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation*, 41:61–89.

A. Fazly, P. Cook, and S. Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

A. Fazly. 2007. *Authomatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph.D. thesis, University of Toronto.

G. Karimi-Doostan. 2011. Separability of light verb constructions in Persian. *Journal of Studia Linguistica*, 65:70–95.

S. Karimi. 1997. Persian complex verbs idiomatic or compositional. *LEXICOLOGY-BERLIN-*, 3(1):273–318.

P. Khanlari. 1973. Tarikh-e zaban-e Farsi [A history of the Persian language]. *Bonyâd-e Farhang*.

N. Mansoory and M. Bijankhan. 2008. The possible effects of Persian light verb constructions on Persian WordNet. In *Proceedings of the Forth Global WordNet Conference (GWC 2008)*. University of Szeged, Department of Informatics.

D. McCarthy, B. Keller, and J. Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

K. Megerdoomian. 2004. A semantic template for light verb constructions. In *Proceedings of the First Workshop on Persian Language and Computers. Tehran University, Iran. May*, pages 25–26.

R. Řehůřek and P. Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

M. Rouhizadeh, A. Yarmohammadi, and M. Shamsfard. 2010. Developing the Persian WordNet of verbs: Issues of compound verbs and building the editor. In *Proceedings of* 5th *Global WordNet Conference*.

S. Stevenson, A. Fazly, and R. North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain*.

A. Villavicencio. 2003. Verb-particle constructions and lexical resources. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 57–64.