

Machine Translation of Persian Complex Predicates

Jan W. Amtrup
Kofax Image Products
5465 Morehouse Dr, Ste. 140
San Diego, CA 92121
(858) 362-3008
jan_amtrup@kofax.com

Karine Megerdooomian
The MITRE Corporation
7525 Colshire Drive
McLean, Virginia 22102
(703) 983-2514
karine@mitre.org

Complex Verbal Predicates (CPs) in Persian consist of a light verb combined with another nonverbal constituent, such as a noun, adjective, or adverb. The challenge of these constructions for machine translation (MT) is that they display both lexical and phrasal properties. They behave as lexical items in that the meaning of the construction is often non-compositional, but they also show phrasal properties as they can participate in syntactic processes; for instance, the constituents can be separated in an input sentence by intervening elements. In Persian, this problem is particularly severe, since most verbs are formed as complex constructions. Thus in many cases, Persian CPs present a translation divergence since they are translated as a single word in the target language (1).

Traditionally, Persian translation systems have treated CPs as lexical items that are explicitly listed in the lexicon as compounds. The problem with this approach is that translation fails as soon as the CP elements are separated, e.g., by a direct object or clitic, modifiers, or relative clauses (2). A slightly more robust representation of CPs within an automatic translation system consists of incorporating their structure inside a complex lexicon, which would treat the preverbal element as “subcategorized” by the light verb, listing the necessary syntax and translation elements with it. However, given that Persian CPs are semi-productive, in particular in the formation of loan words, and that the semantic content can be described in a more regular fashion (cf. Megerdooomian 2002, Folli et al. 2004), a more modular, combinatorial process taking advantage of recent linguistic advances in the study of CPs seems promising. In this approach, we treat the light verb as providing event and aspect information to the complex verbal construction, which then unifies with the information provided by the preverbal element. Hence the verb *bâz kardân* “to open” can be represented as a decomposed change of state verb as shown in the syntactic-semantic template in (3). In the target language, the corresponding lexical material (even if it appears in the form of a single word and not of a complex verb) leads to an identical syntactic-semantic representation, facilitating generation.

This combinatorial analysis of verbs for Persian based on their lexical conceptual structures extends those proposed by Fong et al. (2000) for English and Fujita et al (2004) for Japanese. In this system, the syntactic-semantic template acts as an interlingua representation, allowing for translation between different language pairs. The unification-based, modular analysis of Persian CPs uses several levels of representation: A Template Lexicon which contains the underlying and universal semantic and syntactic templates, and a Vocabulary which includes the words of the particular languages and maps them to the subparts of the semantic-syntactic templates. The preverbs and light verbs are combined during syntactic processing forming the verbal templates provided in the Template Lexicon.

This paper presents a new approach to modeling Persian CPs for MT applications based on a semantic template analysis and compositional formation of the verbal predicates. In this system, there is no need to build a separate lexicon for each language pair and the system can also capture previously unseen constructions, such as the ones formed with loanwords.

Examples

- (1) *jaru zædæn* broom hit 'to sweep'
qosse xordæn worry eat 'to worry'
baz šodæn open become 'to open'
fekr kærdæn thought do 'to think'

- (2) a. *æks-ha ra nešan-æš dad-im*
picture-PL OM sign-CLIT.3SG gave-1PL
'We showed him the pictures.'
b. *qeymæt-e næft æfzayeš-e šædid-i yaft*
price-EZ oil increase-EZ intense-IE found.3SG
'The price of oil increased intensely.'

- (3) [x CAUSE y BECOME <open>]

References

- Folli, R., H. Harley and S. Karimi, (2004) Determinants of Event Types on Persian Complex Predicates, *Lingua* 115: 1365–1401.
- Fong, Sandiway, Christiane Fellbaum and David Lebeaux. 2000. Semantic Templates and Transitivity Alternations in the Lexicon. In Proceedings of TALN 2000, Lausanne, 16-18 October.
- Fujita, Atsushi, Kentaro Furihata, Kentaro Inui, Yuji Matsumoto and Koichi Takeuchi, 2004. Paraphrasing of Japanese Light-verb Constructions Based on Lexical Conceptual Structure. ACL Workshop on Multiword Expressions, Barcelona, Spain, 2004.
- Megerdooonian, Karine. 2002. Beyond Words and Phrases: A Unified Theory of Predicate Composition. Doctoral dissertation, University of Southern California.