

Acquisition of Multiword Lexical Units for FrameNet

FrameNet [1] is a well-known resource for modeling the predicate argument structure of words and organizing them in situation-specific frames and semantic roles (i.e., frame elements). Its interesting formalism to represent the semantics of multiword expressions (MWEs) is often overlooked [2].

FrameNet can represent the relation between constituents of MWEs. The following example from [2] illustrates this: *storage container* and *bread container* evoke the *Container* frame. Roles of this frame are the *Material* of the container, its *Contents*, *Size*, or *Function*. For *storage container*, *storage* fills the *Function* role, while for *bread container*, *bread* fills the *Contents* role (Fig. 1). The FrameNet lexicon model provides the option to annotate *Function* and *Contents* as an “incorporated role” (ICR) for the respective

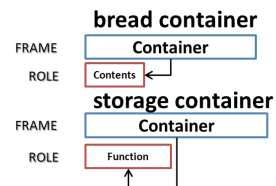


Figure 1: Incorporated roles.

Thus, the implicit relations between the constituents of the MWEs are made explicit. A large FrameNet MWE lexicon can enhance FrameNet-based semantic role labeling (SRL) by a better model for MWEs – see analogous developments integrating MWE detection in parsing [3]. Moreover, the lexicon can be used as information source for the automatic interpretation of MWEs in applications such as information extraction, question answering, or machine translation, for instance by providing features for noun compound interpretation (NCI) [5]. Finally, it provides a basis for further theoretical investigation of MWE semantics.

Unfortunately, the coverage of MWEs in FrameNet 1.5 is low; it contains less than 1,000 multiword entries. This also affects the performance of FrameNet-based SRL [4]. Currently, FrameNet does not make use of its potential to model the relations within MWEs: even though *leather jacket* does occur in the FrameNet example sentences for the *Clothing* frame with the desired incorporated role (*Material*), it does not receive a separate lexical entry.

To close this gap, and to make full use of FrameNet’s potential, an automatic process for the acquisition of MWE lexical units and MWE semantics is desired. Such an automatic approach needs to be based on solid theoretical foundations. Therefore, we present an analysis of the current state of MWEs in FrameNet. Then, we focus on the acquisition of MWE semantics, specifically of ICRs, which, to our knowledge, has not been addressed before. We present a new approach to bootstrap the ICRs of MWEs in FrameNet by annotating their paraphrases with semantic roles, for instance *container that contains bread* for *bread container*. The semantic dependencies between the verb *contains* that evokes the *Container* frame and *bread*, that fills the *Contents* role, mirror the relations between the constituents in *bread container* (Fig. 2). Thus, we can extract the incorporated arguments from the explicit role annotations on the paraphrases. Our approach is related to the work on NCI using paraphrases [6], but is not restricted to compounds and applicable in a multilingual setting. For lexical acquisition of MWEs, previous work on lexical acquisition for FrameNet, for instance using distributional methods [7], can be adapted to MWEs. Our contributions are (i) analyzing the state of MWEs in FrameNet, and (ii) a preliminary evaluation and discussion of the proposed method for ICR detection on MWEs.



Figure 2: Role annotation on paraphrase.

References

- [1] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 86–90, Montreal, Canada, 1998.
- [2] N. Calzolari, C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli. Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain, 2002.
- [3] M. Constant, A. Sigogne, and P. Watrin. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 204–212, Jeju Island, Korea, 2012.
- [4] K. Erk and S. Padó. SHALMANESER – A Toolchain For Shallow Semantic Parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 527–532, Genoa, Italy, 2006.
- [5] S. N. Kim and T. Baldwin. Interpreting Semantic Relations in Noun Compounds via Verb Semantics. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 491–498, Sydney, Australia, 2006.
- [6] P. Nakov. Paraphrasing Verbs for Noun Compound Interpretation. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 46–49, Marrakech, Morocco, 2008.
- [7] M. Pennacchiotti, D. De Cao, R. Basili, D. Croce, and M. Roth. Automatic induction of FrameNet lexical units. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 457–465, Honolulu, Hawaii, 2008.