

Bridging the gap between computable and expressive event representations in Social Media

Darina Benikova

Language Technology Lab
University of Duisburg-Essen
Duisburg, Germany
darina.benikova@uni-due.de

Torsten Zesch

Language Technology Lab
University of Duisburg-Essen
Duisburg, Germany
torsten.zesch@uni-due.de

Abstract

An important goal in text understanding is making sense of events. However, there is a gap between computable representations on the one hand and expressive representations on the other hand. We aim to bridge this gap by inducing distributional semantic clusters as labels in a frame structural representation.

1 Introduction

We experience events in our everyday life, through witnessing, reading, hearing, and seeing what is happening. However, representing events computationally is still an unsolved challenge even when dealing with well-edited text. As soon as non-standard text varieties such as Social Media are targeted, representations need to be robust against alternative spellings, information compression, and neologisms.

The aim of our envisioned project is bridging the gap between *argument-level representations* which are robustly computable, but less expressive and *frame-level representations* which are highly expressive, but not robustly computable. The distinction and the gap between the two main representation types is presented in Figure 1. On the argument-level, the event *give* and all its arguments are identified, whereas on the frame-level additional semantic role labels are assigned.

We envision a representation that enables operations such as equivalence, entailment, and contradiction. In this paper, we will focus on the equivalence operation due to space constraints. These operations are not only necessary to compress the amount of

information, which is especially important in high-volume, high redundancy Social Media posts, but also for other tasks such as to analyze and understand events efficiently.

We plan to achieve building a robustly computable and expressive representation that is suited to perform the discussed operations by using Social Media domain specific clusters and topic labeling methods for the frame-labeling. We intend to evaluate the validity of our representation and approach extrinsically and application-based.

2 Types of representations

We distinguish between representations on two levels: (i) argument-level, which can be robustly implemented more easily, and (ii) frame-level, which is highly expressive.

Argument-level As shown in Figure 1, most argument representations consist of an event trigger, which is mostly a verb, and its corresponding arguments (Banarescu et al., 2013; Kingsbury and Palmer, 2003). Argument-level representations based on Social Media posts are used in applications such as e.g. creating event calendars for concerts and festivals (Becker et al., 2012) or creating overviews of important events on Twitter (Ritter et al., 2012).

Frame-level On this level, events are represented as frame structures such as proposed by Fillmore (1976) that built upon the argument-level, i.e. the arguments are labeled with semantic roles.

A well-known frame semantic tagger is SEMAFOR (Das, 2014).

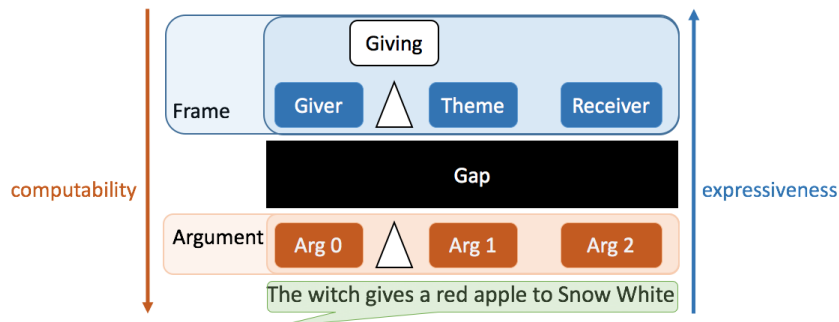


Figure 1: Representations of an exemplary event on argument and frame-level

3 Challenges

The main challenge is to bridge the gap between argument and frame-level representation.

3.1 Performance of operations

Our goal is to develop a representation that is both computable even on noisy Social Media text and expressive enough to support all required operations like equivalence. A semantically equivalent sentence for our exemplary sentence “The witch gave an apple to Snow White” would be “Snow White received an apple from the witch.”, as *receive* is the antonym of *give* and the roles of *Giver* and *Receiver* are inverted.

On the argument-level, it remains a hard problem to establish the equivalence between the two sentences, while that would be easy on the frame-level. However, getting to the frame-level is an equally hard problem and frame representations suffer from low coverage (Palmer and Sporleder, 2010).

3.2 Coverage

Palmer and Sporleder (2010) categorized and evaluated the coverage gaps in FrameNet (Baker et al., 2003). Coverage, whether of undefined units, lemmas, or senses, is of special importance when dealing with non-standard text that contains spelling variations and neologisms that need to be dealt with.

In our opinion, the lack of undefined units is an especially problematic issue in Social Media texts. Furthermore, it may contain innovative, informal or incomplete use of frames, due to space restrictions such as presented by Twitter. Also by cause of space restrictions, which lead to a lack of context, and considering the variety of topics that is addressed in So-

cial Media, it is more challenging to find a fitting frame out of an existing frame repository (Ritter et al., 2012; Li and Ji, 2016).

Giuglea and Moschitti (2006) and Mùjdricza-Mayd et al. (2016) tried to bridge the gap by combing repositories on frame and argument level and representing them based on Intersective Levin Classes (ILC) (Kipper et al., 2006). ILC, which are used in VerbNet (Kipper et al., 2006), are more fine-grained than classic Levin verb classes, formed according to alternations of the grammatical expression of their arguments (Levin, 1993). Classic Levin verb classes were used for measuring semantic evidence between verbs (Baker and Ruppenhofer, 2002).

However, these approaches also have to deal with coverage problems due to their reliance on manually crafted frame repositories.

4 Approach

According to Modi et al. (2012) frame semantic parsing conceptually consists of 4 stages:

1. Identification of frame-evoking elements
2. Identification of their arguments
3. Labeling of frames
4. Labeling of roles

We summarize these tasks in groups of two, namely *identification* and *labeling*, and discuss our approach towards them in the following subsections.

4.1 Identification of frame-evoking elements and their arguments

We regard the first two tasks as tasks of the argument-level, which we plan to solve with part-of-speech tagging and dependency parsing, by extract-

ing all main verbs to solve the first task and considering all its noun dependencies as arguments in the second task. This is similar to the approach of Modi et al. (2012).

4.2 Labeling of predicates and their arguments

Like Modi et al. (2012), we focus on the last two tasks, which we regard as tasks of the frame-level. We observe this task under the aspect of fitting the realization of operation tasks as discussed earlier. As we only regard predicate frames and their arguments for the role labeling, we will use *predicate* as a term for the unlabeled form of *frame* and *argument* as the unlabeled form of *role*.

Pre-defined frame labels There have been attempts to bridge the gap on Social Media texts by projecting ontological information in the form of computed *event types* on the event trigger on the argument-level (Ritter et al., 2012; Li et al., 2010; Li and Ji, 2016) in order to solve the task of frame labeling. However, according to Ritter et al. (2012) the automatic mapping of pre-defined event types is insufficient for providing semantically meaningful information on the event.

We aim to augment those approaches by inducing frame-like structures based on distributional semantics. Moreover, we want to use similarity clusters for the labeling of arguments in frames. We seek to compute the argument labels by the use of supersense tagging, similarly to the approach presented by Coppola et al. (2009). They successfully used the WordNet supersense labels (Miller, 1995) for verbs and nouns as a pre-processing step for the automatic labeling of frames and their arguments.

Approaches using Levin classes, ILC, or WordNet supersenses tackle the same tasks, namely labeling the frame and their corresponding roles. However, all of these suffer from the discussed coverage problem.

Clusters as labels To circumvent the coverage issue, there have been approaches using clusters similarly to frame labels. Directly labeling predicates and their arguments has been performed by Modi et al. (2012), who iteratively clustered verbal frames with their arguments.

As our main goal is to perform operations on event representations, we do not need human-

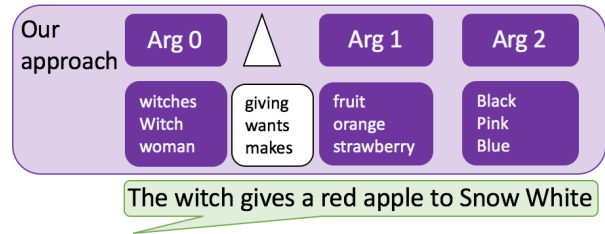


Figure 2: Representations of our approach to bridge the gap

readable frames as proposed by FrameNet, but a level that is semantically equivalent to it, thus our first goal is to compute domain specific clusters for the labeling.

In contrast to Modi et al. (2012), we plan to cluster the verbal predicates and the arguments separately. Although this might seem less intuitive, we believe that due to the difficulties with Social Media data, the structures of full frames are less repetitive and are more difficult to cluster. Thus, by dividing the two tasks of predicate and argument clustering, we hope to achieve better results in our setting.

Furthermore, in order to deal with the issues of the previously discussed peculiarities of the Social Media domain, we plan to train clusters on large amounts of Tweets.

An example of our envisioned representation is shown in Figure 2, which was produced using the Twitter Bigram model of JoBimViz (Ruppert et al., 2015). Figure 2 shows the clustering for finding the correct sense in the labeling task, for both the predicate and its arguments. As the example shows, this representation has some flaws that need to be dealt with. It should be mentioned that the model used for this computation is pruned due to performance reasons, which is a cause for some of the flaws.

For example, *Snow White* is not recognized as a Named Entity or a multi-word expression. To deal with the issue of the false Named Entity representation of *Snow White* presented in the exemplary representation, we plan to experiment with multi-word or Named Entity recognizers. Thus, we plan to train a similar model on a larger set of Tweets, without pruning due to performance reasons.

The main flaw is that the wrong sense cluster of *give* is selected. To improve the issues that occurred above, we plan to use soft-clustering for the step of finding the correct sense cluster. Allowing soft

classes not only facilitates disambiguation (Riedl, 2016), but may also be helpful when identifying the argument role in the frame and thus allow for the previously described operations of equality. By providing only one cluster per predicate Modi et al. (2012) put the task of disambiguation aside, which we want to tackle as mentioned above.

Furthermore, aiming at representations that are suited for operations such as equality, the known problems of antonyms being in the same cluster needs to be solved. Similarly to Lobanova et al. (2010), who automatically extracted antonyms in text, we plan to solve this issue with a pattern-based approach.

Topic-clustered labels After succeeding in the clustering task, we plan to experiment with human-readable frame clusters. In contrast to using pre-defined WordNet supersenses and mapping these to frames, we want to solve the task of finding labels for the clusters by using supersenses computed from domain-specific clusters to directly label the frames and their arguments.

Our hypothesis is that by using more and soft clusters for the supersense tagging, the role labels of the event arguments become semantically richer, because more specific semantic information on the arguments and their context in the event is encoded.

Thus, we plan to use the supersense tagging by using an LDA extension, in which a combination of context and language features is used, as described by Riedl (2016).

5 Evaluation plan

We plan to evaluate our approach in an extrinsic, application-based way on a manual gold standard containing *event paraphrases*. In order to test how well our approach performs in comparison to state-of-the-art approaches of both *argument* and *frame representations*, such as Das (2014) or Li and Ji (2016) in the task of equivalence computation, we will compare the results of all approaches.

For this purpose, we plan to develop a dataset that is similar to Roth and Frank (2012), but tailored to the Social Media domain. They produced a corpus of alignments between semantically similar predicates and their arguments from news texts on the same event.

6 Summary

In this paper we present our vision on a new event representation that enables the use of operations such as equivalence. We plan to use pre-processing to get the predicates and their arguments. The main focus of the work will be using sense clustering methods on domain-specific text and to apply these clusters on text. We plan to evaluate this application in an extrinsic, application-based way.

Further on, we plan to tackle tasks such as: topic-model based frame labeling on the computed clusters; pattern-based antonym detection in the clusters for enabling the operation of contradiction and improve the task of equivalence; and experiment with Named Entity and multiword recognizers in order to improve the results in argument recognition.

7 Acknowledgements

This work is supported by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

We also kindly thank the EMNLP for granting us a travel support.

References

- Collin F. Baker and Josef Ruppenhofer. 2002. FrameNet's frames vs. Levin's verb classes. In *Proceedings of the 28th annual meeting of the Berkeley Linguistics Society*, pages 27–38.
- Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The structure of the framenet database. *International Journal of Lexicography*, 16(3):281–296.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.
- Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. 2012. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 533–542, Seattle, WA, USA.
- Bonaventura Coppola, Aldo Gangemi, Alfio Gliozzo, Davide Picca, and Valentina Presutti. 2009. Frame detection over the Semantic Web. In *European Semantic Web Conference*, pages 126–142. Springer.

- Dipanjan Das. 2014. Statistical models for frame-semantic parsing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore*, pages 26–29, Baltimore, MD, USA.
- Charles J. Fillmore. 1976. Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences*, pages 20–33.
- Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic Role Labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 929–936, Sydney, Australia.
- Paul Kingsbury and Martha Palmer. 2003. PropBank: the Next Level of TreeBank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 105–116, Vaxjo, Sweden.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of Language Resources and Evaluation*, pages 1027–1032, Genoa, Italy.
- Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.
- Hao Li and Heng Ji. 2016. Cross-genre Event Extraction with Knowledge Enrichment. In *The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1162, San Diego, CA, USA.
- Hao Li, Xiang Li, Heng Ji, and Yuval Marton. 2010. Domain-Independent Novel Event Discovery and Semi-Automatic Event Annotation. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 233–242, Sendai, Japan.
- Anna Lobanova, Tom Van der Kleij, and Jennifer Spender. 2010. Defining antonymy: A corpus-based study of opposites by lexico-syntactic patterns. *International Journal of Lexicography*, 23(1):19–53.
- George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. 2012. Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 1–7, Montral, Canada. Association for Computational Linguistics.
- Eva Mùjdricza-Mayd, Silvana Hartmann, Iryna Gurevych, and Anette Frank. 2016. Combining semantic annotation of word sense & semantic roles: A novel annotation scheme for verbnet roles on german language data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3031–3038, Portorož, Slovenia, May.
- Alexis Palmer and Caroline Sporleder. 2010. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Proceedings of the 23rd international conference on computational linguistics*, pages 928–936, Uppsala, Sweden.
- Martin Riedl. 2016. *Unsupervised Methods for Learning and Using Semantics of Natural Language*. Ph.D. thesis, TU Darmstadt.
- Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112, Beijing, China.
- Michael Roth and Anette Frank. 2012. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 218–227, Montral, Canada.
- Eugen Ruppert, Manuel Kaufmann, Martin Riedl, and Chris Biemann. 2015. Jobimviz: A web-based visualization for graph-based distributional semantic models. In *The Annual Meeting of the Association for Computational Linguistics (ACL) System Demonstrations*, pages 103–108.