Enriching Frame Representations with Distributionally Induced Senses

Stefano Faralli†, Alexander Panchenko‡, Chris Biemann‡, Simone Paolo Ponzetto†
†Data and Web Science Group, School of Business Informatics and Mathematics, Universität Mannheim, Germany
‡ Language Technology Group, Department of Informatics, Universität Hamburg, Germany
{stefano,simone}@informatik.uni-mannheim.de
{panchenko,biemann}@informatik.uni-hamburg.de

Abstract

We introduce a new lexical resource that enriches the Framester knowledge graph, which links Framnet, WordNet, VerbNet and other resources, with semantic features from text corpora. These features are extracted from distributionally induced sense inventories and subsequently linked to the manually-constructed frame representations to boost the performance of frame disambiguation in context.

Keywords: distributional semantics, linked open data, frame semantics.

1. Introduction

Recent years have witnessed an impressive amount of work on the automatic construction of wide-coverage knowledge resources. Web-scale information extraction systems like NELL (Carlson et al., 2010) or Knowledge Vault (Dong et al., 2014) can acquire massive amounts of machine-readable knowledge from the Web, whereas projects like DBpedia (Bizer et al., 2009), YAGO (Rebele et al., 2016) or BabelNet (Navigli and Ponzetto, 2012) have turned collaboratively-generated content into large knowledge bases. However, all of these resources are entity-centric in that they are primarily built around the notion of *entities*, as either provided by an external resource (e.g., Wikipedia pages) or automatically discovered from text (e.g., by clustering entity mentions). The entities are most commonly represented by nouns, noun phrases, and named entities. Entities lie at the heart of the Semantic Web and are central to enable semantic technologies on a large scale (Dietz et al., 2017). Besides, they also provide the foundation for more complex semantic representations like event templates or frames (Fillmore, 1968), which are in the focus of our interest in this paper.

Recent work looked at ways to populate the Linked Open Data (LOD) cloud with wide-coverage information about semantic frames on the basis of Framester (Gangemi et al., 2016a), a frame-centric resource that is meant to act as a hub between several other linguistic resources that are already part of the LOD cloud. In parallel, researchers looked at ways to combine knowledge graphs of this kind with distributional semantics to include human-readable meaning representations based on semantic vector spaces within the Semantic Web ecosystem (Faralli et al., 2016). Consequently, in this paper we bring these two lines of research together and present a new resource that combines a wide-coverage symbolic repository of frames with human-readable distributional semantic representations from text.

In Table 3, we show an excerpt from a distributional-based sense inventory, a proto-conceptualization (PCZ) proposed by Faralli et al. (2016) and further developed by Biemann et al. (2018). The senses in this resource are induced automatically from text via clustering of graphs of semantically related words and linked to BabelNet and WordNet lexical resources. In particular, in the resource, each sense (e.g., *read#VB#1*) is defined by its related senses (*quote#VB#0*, *recite#VB#0*, and so on) which have been observed to co-occur with the sense. Sense inventories with related senses as well as other features, such as weighted hypernymy senses and context clues, can provide rich and human-readable knowledge to disambiguate word meanings in context, e.g., the meaning for the verb *to read* as 'read aloud' versus 'make sense' as in:

When we *hear* someone READ a text, our *understanding* of what we *hear* is usually spontaneous – the rules by which we *interpret* meaning.

Here, the reference to the 'Reading_aloud' frame from FrameNet[1] could be triggered on the basis of the occurrences of the verbs in the sentence: *hear*, *understanding* and *interpret*. Such connections, in turn, could be provided by a hybrid resource where distributional representations from text have been explicitly linked to semantic knowledge repositories, since hybrid resources of this kind have been shown in the past to improve performance on lexical understanding (Panchenko et al., 2017) as well as taxonomy learning and cleaning (Faralli et al., 2017).

In this paper, we bridge the gap between distributional and frame semantics by linking distributional semantic representations to Framester, a knowledge graph that acts as a hub between resources like FrameNet, BabelNet and DBpedia, among others. As a result of this, we introduce a new lexical resource that enriches the Framester knowledge graph with distributional features extracted from text, and show how this hybrid resource yields better results on the task of recognizing frames in running text.

Joining distributional and frame semantics builds upon and extends our framework for combining symbolic and statistical meaning representations (Biemann et al., 2018): the main objective of this line of research is to 'join forces' across heterogeneous knowledge and semantic models in order to mutually enrich them and combine the strengths

[1] https://framenet.icsi.berkeley.edu/

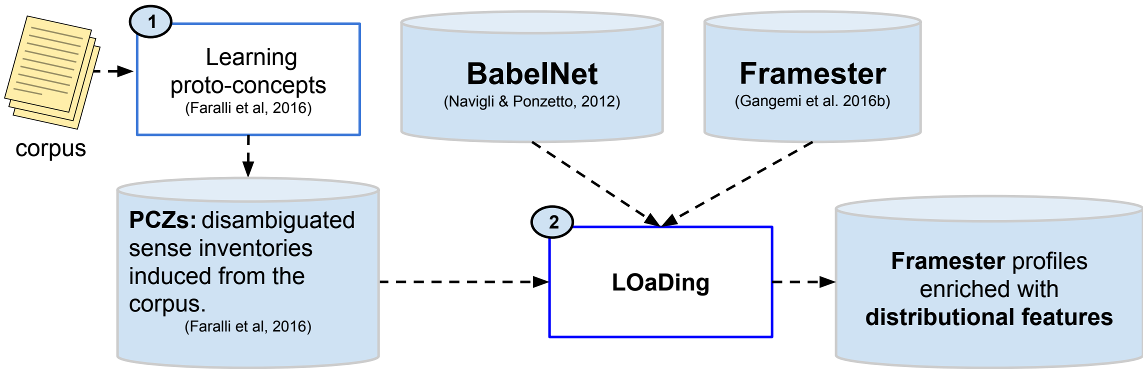

Figure 1: The overall resource construction workflow of our approach: distributionally induced sense representations extracted from a corpus (PCZ) are used to enrich the Framester symbolic knowledge graph.

BabelNet Synset ID	English Word Senses	Gloss
bn:00085007v	quote, cite	'To repeat someone's exact words.', 'To quote; to repeat, as a passage from a book, or the words of another.'
bn:00090740v	observe, mention, note, remark	'She observed that his presentation took up too much time.', 'They noted that it was a fine day to go sailing.'

Table 1: Excerpt of two BabelNet synsets.

Frame ID	Related BabelNet Synsets
Evidence	bn:00085007v, bn:00084633v, ...
Telling	bn:00085007v, bn:00083488v, ...
Communication	bn:00085007v, bn:00090740v, ...

Table 2: Excerpt of the Framester TransX profile.

of the lexicographic tradition that describes linguistic information manually with the coverage and versatility of the corpus and data-driven approaches that derive meaning representations directly from the data. Bringing together the 'best of both worlds' has the potential to combine the benefits of wide-coverage symbolic (Gangemi et al., 2016b) and statistical (Chen et al., 2014) semantics for frame parsing, which, in turn, can be exploited for many different applications ranging from sentiment analysis (Recupero et al., 2015) all the way to content-based recommendations (De Clercq et al., 2014).

The contributions of this paper are the following ones:

1. We present the *LOaDing* lexical resource, an extension of Framester that adds distributional-based features (Faralli et al., 2016; Biemann et al., 2018) to frame representations (see Section 2.);

2. We evaluate the resource in the task of Word Frame Disambiguation (WFD), i.e., on the identification of frames in context (Gangemi et al., 2016b), and show significant improvements over the original Framester frame representations (see Section 3.).

Both resources and software produced in this work are available under a CC-BY 4.0 license.[2]

[2] http://web.informatik.uni-mannheim.de/joint/

2. Enriching Framester with Symbolic Distributional Sense Representations

In Figure 1 we show the diagram describing the workflow used to add distributional features to the Framester profiles to enable a better word frame disambiguation.

2.1. Resources and Datasets

Our approach makes use of three linguistic resources:

BabelNet (Navigli and Ponzetto, 2012): a multilingual encyclopedic dictionary that connects concepts and named entities in a very large network of semantic relations (see Table 1).

Framester (Gangemi et al., 2016a): a linguistic LOD hub that provides links from FrameNet's frames to semantically related BabelNet senses (see Table 2). The current version consists of six different profiles: *Base* contains only manually curated links, whereas the other five (*DirectX*, *FrameBase*, *Fprofile*, *TransX* and *XWFN*) are automatically built extensions or subsets of the *Base* profile.

Proto-Conceptualizations (PCZs) (Faralli et al., 2016): a fully disambiguated sense inventory automatically induced from text, providing human-readable distributional semantic representations. In this resource, each sense is provided with (see Table 3): a) a weighted list of semantically-related and hypernymy senses; b) links to existing knowledge bases (i.e., BabelNet); c) context clues used to disambiguate the senses in the context.

2.2. Combination of Resources

The goal of our method is a combination of the three resources mentioned in the previous section as illustrated

PCZ Sense ID	BabelNet Synset ID	Related Senses	Context Clues
quote#VB#0	bn:00085007v	cite#NN#1[1.0], interview#VB#0[0.8], mention#VB#1[0.7], publish#VB#0[0.6], review#VB#0[0.6], ...	in#IN#pcomp[24799.3], Register#NP#prep_on[21282.9], Track#NP#-vmod[16808.9], ...
mention#VB#1	bn:00090740v	attest#VB#0[1.0], describe#VB#0[0.8], document#VB#0[0.7], quote#VB#0[0.6], ...	Register#NP#prep_on[45477.7], Track#NP#-vmod[35850.4], say#VB#prepc_as[17041.8], ...

Table 3: Excerpt from a proto-conceptualization (PCZ) of two distributional-based senses. Numbers behind related senses are normalized similarity scores, numbers behind context clues are association scores.

Frame ID	Related BabelNet Synsets	Related PCZ Senses
Evidence	bn:00084633v[29.0], bn:00085007v[13.0], ...	quote#VB#0, ...
Telling	bn:00085007v[12.0], bn:00083488v[8.0], ...	quote#VB#0, ...
Communication	bn:00090740v[18.0], bn:00085007v[15.0], ...	mention#VB#1, quote#VB#0, ...

Table 4: Excerpt of a *LOaDing* TransX profile. Numbers behind related synsets bs are the distributional semantics-based weights $w(F, bs)$ we assign in Step 2.4 of our LOaDing approach.

in Table 4. This is made possible by using BabelNet as a pivot, since Framester provides links from FrameNet to BabelNet (Table 2) while our PCZs are also linked to BabelNet (Table 3). As a first result of such a combination we created *LOaDing*, an extension of Framester's profiles where we provide weights for each frame-related BabelNet synset. We show in Table 4 an excerpt of the resulting extended Framester's TransX profile shown in Table 2.

The enriched frame representations are built in two main steps as illustrated in Figure 1:

Step 1: Learning a proto-conceptualization (PCZ). We apply the methodology from Biemann et al. (2018), to produce disambiguated sense inventories with entries for nouns and verbs.

Step 2: LOaDing a PCZ into Framester. For each frame F, e.g. *Communication*, and related BabelNet sysnet bs, e.g. *bn:00085007v*, we assign a weight $w(F, bs)$. In order to compute the weight, we create three bag-of-word representations for each of the three resources listed in the previous section:

1. $B_{bn} = BoW(bs)$ to represent bs including the counts for all the content words from the synset word senses and the glosses;

2. From the PCZ we collect all the entries $PCZ(bs)$ matching bs, for instance $PCZ(\textit{bn:00085007v}) = \{\textit{quote\#VB\#0}, \dots\}$, and create a bag of words $B_z = BoW(PCZ(bs))$ by collecting all the weights of content words from the senses $s \in PCZ(bs)$ and the corresponding related senses.

3. We create a BoW from B_f by collecting all the content words from the description of the frame F.

4. Finally, we compute a relatedness score between a frame F and a BabelNet synset bs:

$$w(F, bs) = \sum_{w \in I} B_z.c(w) \times (B_f(F).c(w) + B_{bn}.c(w)),$$

where $I = B_z(bs) \cap (B_{bn}(bs) \cup B_f(F))$ and $A.c(w)$ equals the number of occurrences of the word w in A. For instance, with respect to the excerpts of Tables 3 and 4 we obtain $w(\textit{Communication}, \textit{bn:00085007v}) = 15.0$.

3. Using the Enriched Representations for Word Frame Disambiguation

We evaluate our extensions of Framester profiles following the experimental setting of Gangemi et al. (2016b), and compare the extended and the original profiles in a task of Word Frame Disambiguation (WFD).

3.1. Dataset: FrameNet Full Text Documents

To create a silver standard we processed all 108 documents from the FrameNet 1.7 dataset (Baker et al., 1998) with BabelFy (Moro et al., 2014)[3]. By combining the original frame annotations with the automatically generated entity links we collected a total of 81,706 annotations, which we use in our experimental setting as a silver standard.

3.2. Word Frame Disambiguation

Following the WFD approach described in Gangemi et al. (2016b) we implemented a simple word frame disambiguator, where for each provided annotation in our silver standard we try to predict a frame label only on the basis of the BabelNet synsets generated through BabelFy. In order to provide the most suitable frame label F for the provided BabelNet synset label bs:

1. Creating the set of candidate frames $CF(bs)$, by collecting all the frame IDs for which bs is a semantically related BabelNet synsets (e.g., $CF(\textit{bn:00085007v})= \{\textit{Evidence}, \textit{Telling}, \textit{Communication}\}$ see Table 4);

2. Ranking all the candidate frames $f \in CF(bs)$ by computing the following scoring functions:

[3] http://babelfy.org

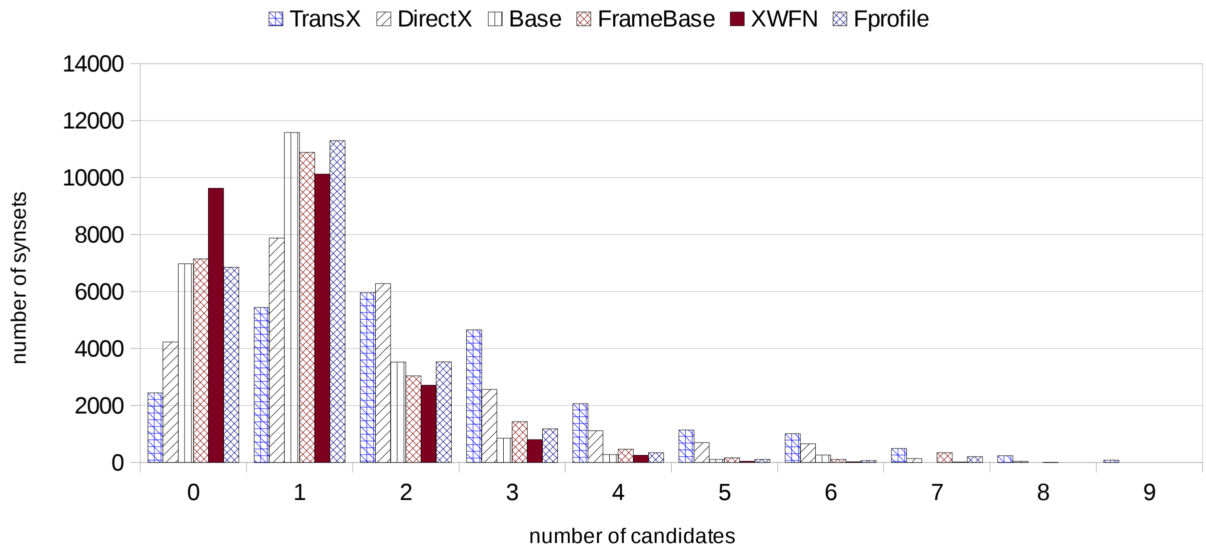

Figure 2: The distribution of synset counts per number of candidate frames (step 1 of our WFD approach, see Section 3.2.). The y axis counts the synsets bs for which $|CF(bs)| = x$.

	Base	FrameBase	XWFN	TransX	DirectX	Fprofile
oracle	67.60	67.74	57.85	77.82	73.74	68.16
original (cond)	53.13	52.73	45.44	**45.21**	46.30	53.04
wiki-n30-1400k (cond)	53.23	52.82	45.42	44.75	**46.58**	53.46
wiki-n30-1400k (inv)	**55.52**	**55.06**	**46.73**	28.24	45.39	**55.69**

Table 5: F1 performances on Word Frame Disambiguation across profiles and ranking methodologies.

- the score is equal to the conditional probability of a frame given a synset (*cond*):

$$score_{cond}(f, bs) = \frac{w(f, bs)}{\sum_{b \in relatedSynsets(f)} w(f, b)};$$

- the score is equal to the number of related synsets divided by the weight of the synsets (*inv*):

$$score_{inv}(f, bs) = \frac{|relatedSynsets(f)|}{w(f, bs)}.$$

  Such a scoring function promotes the candidate frame that relates with the highest number of synsets $|relatedSynsets(f)|$, and it also penalizes the selection of candidates which are triggered by synsets having higher weights, in other words promoting candidates triggered by synsets in the 'long tail';

3. Finally, we select the candidate frame with the highest score.

We experimented with a variety of scoring functions based on weights $w(F, bs)$ and $|relatedSynsets(F)|$, and present only the two best performing ones here.
In Figure 2 we draw the distribution of the counts of synsets across the number of frames that are triggered in the phase of candidate selection (step 1 of our WFD approach), i.e., in the XWFN profile, we counted ten thousand synsets bs for which $|CF(bs)| = 1$.

3.3. Results

In Table 5, we show the resulting word frame disambiguation performances in terms of the F-score and compare across profiles and ranking methodologies. To better understand the limits of our 'silver standard' and the limits of ranking methodologies we introduced a so called 'oracle', which selects the labeled frame if present in the list of selected candidates independently from its position in the ranking. The results indicate that:

- For Base, FrameBase, XWFN and Fprofile profiles, adding distributional-semantic-based features lead to 1 to 2 points improvements of F1 in the WFD task;
- Our approach lowers the performance of DirectX profile and even drastically lowers the performances of the TransX profile;
- Overall, we noticed that the former two profiles are very dense by means of related senses per frame and lead to a potential high recall (see the oracle performances on the two profiles at 73.74%, 77.82% respectively) while introducing more noisy senses;
- Finally, the presence of many frequent and noisy senses let the $score_{inv}$ function outperform other function based on weights $w(F, bs)$ and $|relatedSynsets(F)|$.

To summarize this experiment: when using our induced weighting with the $score_{inv}$ ranking, we achieve a boost on Framester profiles that showed a good performance in the original setting, while losing performance on inferior, more noisier profiles. The overall best performance was reached by our extension of the high-quality, manually created Base profile, followed by our extension to the Fprofile.

3.4. Error analysis

An in-depth analysis of the errors made during the experiments (see Section 3.3.) by our WFD approach (see Section 3.2.) allowed us to identify the following categories of errors:

- *Misaligned annotations:* our silver standard is automatically created by collecting BabelFy annotations for a total of 81,706 annotations from the original 93,358 FrameNet 1.7 annotations;
- *Entity linking errors:* we estimated that around 15% of the 81,706 BabelFy annotations[4] assigned a wrong sense, eventually triggering the selection of a wrong frame;
- *Ranking error:* errors where an inferior ranking leads to bad performance. These are a consistent part of the WFD errors and may be improved with more complex weighting/ranking approaches;
- *Profile errors:* related to a specific profile. In this category we identify errors due to: i) the slightly difference between some frame meaning (e.g., many errors are due to the prediction of the frame *Measure duration* instead of *Calendric unit*); ii) the frequent absence of related senses so that is not possible to trigger any frame, even for correct BabelFy annotations.

4. Conclusions

In this paper, we presented *LOaDing*, a novel lexical resource that connects distributional sense representations induced from text to symbolic frame representations. Hybrid semantic representations have been shown to enable complex semantic tasks like, for instance, end-to-end taxonomy induction (Faralli et al., 2017): they could provide an additional signal to improve state-of-the-art semantic parsing (as illustrated by our running example). As we have shown for word sense disambiguation for nouns (Panchenko et al., 2017), distributional information is able to considerably alleviate the sparsity inherent in knowledge-based methods. Consequently, in future work we would like to study the contribution and the potential of our proto-conceptualizations into more complex frame-centered tasks such as verb frame induction (Vulić et al., 2017). Our vision in the longer term is to exploit hybrid statistical and symbolic approaches to go beyond the vocabulary and relations of entity- and encyclopedic-centric resources to produce novel semantic representations of event templates and frames with a large coverage.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG) within the JOIN-T project.

[4] A sample of 1000 annotations has been manually annotated as correct or wrong by two annotators, and a third annotator provided a final judgment only for the annotations in disagreement.

5. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley Framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Biemann, C., Faralli, S., Panchenko, A., and Ponzetto, S. P. (2018). A framework for enriching lexical semantic resources with distributional semantics. *Natural Language Engineering*, pages 1–48.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – A crystallization point for the web of data. *JWS*, 7(3):154–165.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of AAAI*, pages 1306–1313, Atlanta, GA, USA. AAAI Press.

Chen, Y. N., Wang, W. Y., and Rudnicky, A. I. (2014). Leveraging frame semantics and distributional semantics for unsupervised semantic slot induction in spoken dialogue systems. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, pages 584–589, South Lake Tahoe, NV, USA. IEEE.

De Clercq, O., Schuhmacher, M., Hoste, V., and Ponzetto, S. P. (2014). Exploiting FrameNet for content-based book recommendation. In *Proceedings of CBRecSys Workshop at RecSys*, pages 14–20, Foster City, CA, USA. ACM.

Dietz, L., Kotov, A., and Meij, E. (2017). Utilizing knowledge graphs in text-centric information retrieval. In *Proceedings of WSDM*, pages 815–816. ACM.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of KDD*, pages 601–610, New York, NY, USA. ACM.

Faralli, S., Panchenko, A., Biemann, C., and Ponzetto, S. P. (2016). Linked disambiguated distributional semantic networks. In *Proceedings of ISWC*, pages 56–64, Kobe, Japan. Springer International.

Faralli, S., Panchenko, A., Biemann, C., and Ponzetto, S. P. (2017). The ContrastMedium algorithm: Taxonomy induction from noisy knowledge graphs with just a few links. In *Proceedings of EACL*, pages 590–600, Valencia, Spain. Association for Computational Linguistics.

Fillmore, C. J. (1968). The case for case. In *Universals in Linguistic Theory*, pages 2–88. Holt, Rinehart and Winston, New York, NY, USA.

Gangemi, A., Alam, M., Asprino, L., Presutti, V., and Recupero, D. R. (2016a). Framester: A wide coverage linguistic linked data hub. In *Proceedings of EKAW*, pages 239–254, Bologna, Italy.

Gangemi, A., Alam, M., and Presutti, V. (2016b). Word frame disambiguation: Evaluating linguistic linked data on frame detection. In *Proceedings of LD4IE Workshop* *at ISWC*, pages 23–31, Kobe, Japan. Springer International.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *ArtInt.*, 193:217–250.

Panchenko, A., Faralli, S., Ponzetto, S. P., and Biemann, C. (2017). Using linked disambiguated distributional networks for word sense disambiguation. In *Proceedings of SENSE Workshop at EACL*, pages 1–7, Valencia, Spain. Association for Computational Linguistics.

Rebele, T., Suchanek, F. M., Hoffart, J., Biega, J., Kuzey, E., and Weikum, G. (2016). YAGO: A multilingual knowledge base from Wikipedia, Wordnet, and Geonames. In *Proceedings of ISWC*, pages 177–185, Kobe, Japan. Springer International.

Recupero, D. R., Presutti, V., Consoli, S., Gangemi, A., and Nuzzolese, A. G. (2015). Sentilo: Frame-based sentiment analysis. *Cognitive Computation*, 7(2):211–225.

Vulić, I., Mrkšić, N., and Korhonen, A. (2017). Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*, pages 2536–2548, Copenhagen, Denmark. Association for Computational Linguistics.