Semantic Frame Induction from a Real-World Corpus

Anonymous ACL submission

Abstract

Recent studies on semantic frame induction have demonstrated that the emergence of pretrained language models (PLMs) has led to more accurate results. However, most existing studies evaluate the performance using frame resources such as FrameNet, which may not accurately reflect real-world language usage. In this study, we conduct semantic frame induction using the Colossal Clean Crawled Corpus (C4) and assess the applicability of existing frame induction methods to real-world data. Our experimental results demonstrate that existing frame induction methods are effective on real-world data and that frames corresponding to novel concepts can be induced.

1 Introduction

005

011

012

015

017

021

037

041

Frame semantics (Fillmore, 1982) assumes that humans rely on background knowledge derived from experience and world knowledge when interpreting language. Such knowledge is referred to as semantic frames. These frames are evoked by specific words or phrases, referred to as frame-evoking expressions, or lexical units (LUs) in FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016). Semantic frame induction is the task of clustering frame-evoking expressions in context according to the frames they evoke. It constitutes an important step toward the automatic construction of semantic frame resources for specific domains and lowresource languages using large corpora (QasemiZadeh et al., 2019). Recent studies on semantic frame induction (Ribeiro et al., 2019; Anwar et al., 2019; Arefyev et al., 2019; Yamada et al., 2021, 2023) have employed contextualized word embeddings such as BERT (Devlin et al., 2019), and these approaches have outperformed traditional methods (Ustalov et al., 2018; Materna, 2012).

However, despite the goal of constructing realworld semantic resources, most studies evaluate the performance of semantic frame induction based on existing frame resources such as FrameNet, which may not accurately reflect real-world language usage. Specifically, two points can be raised as differences between FrameNet and real-world corpora. First, the frequency distribution of lexical items and their semantic usages in FrameNet differs from that observed in real-world corpora. FrameNet provides both lexicographic annotations, which tag manually selected examples for predefined LUs, and full-text annotations, which tag all frame-evoking expressions in text. However, only 14% of the examples are full-text annotations,¹ limiting its representativeness of real-world language. Second, FrameNet lacks coverage of recent vocabulary and usages. For example, the usage of the verb "stream" meaning "to send or receive sound or video directly over the internet" is not included in FrameNet. Our analysis revealed that 90.2% of verb-related annotations were created in or before 2008, suggesting that the data may be outdated.

042

043

044

047

048

053

054

056

060

061

062

063

064

065

066

067

068

069

070

071

072

073

074

076

077

078

079

Differences in the frequency distribution of word senses across corpora may influence the difficulty of semantic frame induction. Thus, it is unknown to what extent existing frame induction methods are applicable to real-world corpora. Moreover, applying frame induction to more recent and diverse corpora has the potential to uncover novel frames that are not covered in existing frame resources. To explore these issues, this study conducts frame induction using examples extracted from the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020) and analyzes the induced results. A key challenge is that real-world corpora lack gold-standard frame annotations, making direct evaluation difficult. To address this, we propose an evaluation method that indirectly assesses induced clusters by comparing them with FrameNet examples, enabling analysis of their alignment with existing frames and their ability to capture emerging usage.

¹http://framenet.icsi.berkeley.edu/current_status (accessed on May 2024)

087

096

098

100

101

102

103

104

106

107

108

110

111

112

2 Semantic Frame Induction with Deep Metric Learning

In this study, we focus on verbs as frame-evoking expressions and adopt the method proposed by Yamada et al. (2023) for semantic frame induction. Their approach first generates contextualized embeddings for frame-evoking verbs in the examples and then performs clustering to induce semantic frames. It employs two clustering methods: i) onestep clustering that clusters all verb examples at once, and ii) two-step clustering that first clusters examples for each verb individually and then performs clustering across verbs. To reduce the influence of surface-level lexical information, it utilizes masked word embeddings. Specifically, as shown in Equation (1), the final embedding v_{w+m} for a frame-evoking verb is computed as a weighted average of the standard embedding $v_{\rm word}$ and the embedding of the [MASK] token $v_{\rm mask}$ when the verb is replaced with a [MASK] token:

$$v_{w+m} = (1 - \alpha) \cdot v_{\text{word}} + \alpha \cdot v_{\text{mask}}.$$
 (1)

Furthermore, to obtain embeddings that are better suited for semantic frame induction, the contextualized embedding model is fine-tuned using a portion of the annotated examples in FrameNet with deep metric learning (Kaya and Bilge, 2019; Musgrave et al., 2020). During training, the model is optimized so that embeddings of frame-evoking verbs that belong to the same frame are drawn closer together, while those belonging to different frames are pushed farther apart.

3 Experimental Setup and Evaluation

Figure 1 presents an overview of our framework. 113 First, we apply Yamada et al. (2023)'s frame in-114 duction method to examples extracted from the C4 115 corpus. Here, the verb distribution in the frame in-116 duction examples is aligned with that of FrameNet, 117 which serves as the evaluation reference. Next, 118 to assess the validity of the constructed clusters, 119 we perform an evaluation using examples from FrameNet. Specifically, each FrameNet example 121 is mapped to the nearest C4 example in the em-122 bedding space and assigned to the cluster to which 123 that example belongs. We then conduct a quan-125 titative evaluation of the induced frames, treating the FrameNet annotations as ground truth. Finally, 126 we perform a qualitative analysis of the induced 127 clusters, particularly those that are not aligned with any examples in FrameNet. 129



Figure 1: Overview of our framework. In the figure, each \bullet represents an example extracted from the C4 corpus, and its color indicates the cluster to which the example belongs. The symbols \bullet , \blacktriangle , and \star represent examples from FrameNet. Identical symbols indicate that the examples are annotated with the same frame. Arrows pointing to \bullet indicate the corresponding examples in the C4 corpus. The cluster consisting solely of \bullet examples has no FrameNet counterpart and is therefore a candidate for novel frames.

3.1 Extracting Examples from C4

We extract a set of example sentences from the C4 corpus for frame induction. As described above, we evaluate the frame induction results by aligning the FrameNet examples with the examples from the C4 corpus. If the distribution of frame-evoking verbs in the C4 examples differs substantially from that in the FrameNet evaluation set, some FrameNet examples may lack corresponding assignments, potentially compromising the reliability of the evaluation. To mitigate this issue, we extract examples from C4 such that the distribution of frame-evoking expressions is consistent with that of the FrameNet evaluation set.

It should be noted, however, that the distribution of semantic usages for each frame-evoking expression in C4 is unknown and does not match that in FrameNet. Therefore, the extracted examples may include instances that evoke novel frames not covered by FrameNet. For example, consider the frame-evoking verb "stream." In FrameNet, this verb appears only as LUs in the Mass_motion and Fluidic_motion frames. However, in recent 130

131

132

133

146

147

148

149

150

years, stream is more frequently used in a relatively recent sense "to send or receive sound or video directly over the internet." Since the examples of each frame-evoking verb extracted from C4 are randomly sampled, they are assumed to reflect the actual usage distribution. Therefore, it is expected that novel frames corresponding to such recent meanings may be induced.

161

163

164

165

167

168

170

171

172

173

174

175

176

177

178

179

181

183

184

187

190

191

194

195

196

197

198

199

201

3.2 Quantitative and Qualitative Analysis of Induced Frames using FrameNet

Since the examples extracted from C4 are not annotated with frame information, a key challenge is how to evaluate semantic frame induction performed on such data. To address this issue, we conduct an evaluation leveraging FrameNet data, as illustrated in Figure 1.

The motivation for this analysis is as follows. As a premise, 86% of the examples in FrameNet originate from lexicographic annotations, which are carefully curated to reflect prototypical usages of each frame. In contrast, examples extracted from the C4 corpus are not curated in this way and may contain marginal or ambiguous usages. Consequently, clustering C4 examples presents a more challenging task. If, despite this increased difficulty, clustering C4 examples yields frames similar to those induced from FrameNet examples, it would suggest that the frame induction method is robust to real-world data. In such cases, we can assume that mapping each FrameNet example to its most similar C4 example and assigning it to the corresponding cluster should ideally result in clusters that correspond to the frames evoked by the FrameNet examples.

To quantitatively analyze the induced frames, we evaluate the performance of frame induction by comparing the frame annotations in FrameNet with the cluster assignments obtained through the mapping procedure. As evaluation metrics, we use B-cubed F1 (BCF) (Bagga and Baldwin, 1998) and the harmonic mean of Purity and Inverse Purity (PIF) (Zhao and Karypis, 2001).

We also conduct a manual qualitative evaluation of the induced frames. Some clusters are not aligned with any FrameNet examples, and may correspond to frames not covered by FrameNet. Accordingly, we place particular emphasis on analyzing these clusters to investigate whether they represent novel frames.

	#Verbs	#LUs	#Frames	#Instances
Set 1	827	1,255	433	26,835
Set 2	827	1,299	424	27,210
Set 3	827	1,276	436	27,225
All	2,481	3,830	637	81,270

Table 1: Statistics of the FrameNet dataset used in threefold cross-validation.

Clustering	Model	PIF / BCF		
		C4	FrameNet	
One-step	Vanilla	49.7 / 36.3	54.9 / 42.6	
	Triplet	70.9 / 60.8	73.3 / 63.9	
	Softmax	71.4 / 60.0	71.0 / 59.9	
	AdaCos	73.3 / 62.6	73.5 / 62.9	
Two-step	Vanilla	36.5 / 20.9	66.9 / 56.5	
	Triplet	66.0 / 53.6	74.8 / 66.0	
	Softmax	70.2 / 58.9	71.9 / 61.4	
	AdaCos	70.7 / 59.6	75.6 / 66.2	

Table 2: Evaluation results of frame induction with C4 examples. Scores for FrameNet are cited from Yamada et al. (2023).

3.3 Experimental Settings

We conducted experiments using three-fold crossvalidation, in which the FrameNet examples were divided into three subsets by verb serving as training, development, and test data. Table 1 shows the statistics for each split. The training set is used as training data for deep metric learning; the development set is used to determine the weight α in Equation (1), the number of clusters, and the margin for loss functions.

We use the pre-trained BERT model (bert-baseuncased) as our contextualized word embedding model and FrameNet 1.7 (Ruppenhofer et al., 2016) as the frame resource. For clustering, we employ two methods: one-step clustering using agglomerative (group-average) clustering, and two-step clustering, in which X-means clustering (Pelleg and Moore, 2000) is first applied to individual verbs, followed by group-average clustering across verbs. For deep metric learning, we experiment with three loss functions: Triplet (Weinberger and Saul, 2009), Softmax (Liu et al., 2017), and AdaCos (Zhang et al., 2019). We also conduct experiments in a vanilla setting, where we use the pre-trained BERT model without fine-tuning.

4 Experimental Results

Quantitative analysis Table 2 summarizes the quantitative evaluation results of semantic frame

Induced frames	C4 examples (boldface indicates the frame-evoking verb)		
Education_teaching	tutor students in math / can tutor you / trained for working with children		
Violation	violate privacy / contravene those rules. / company has breached the law		
Cause_to_hasten	Do not rush yourself! / should not rush a patient / being hastened		
Media_streaming	stream the video / stream the video / be streamed on 5G.		

Table 3: Examples of induced frames. Their names were manually assigned based on the semantics of the corresponding instances. The top two frames are aligned with many FrameNet examples. The bottom two frames are not aligned with any FrameNet examples, representing potentially novel frames.

induction.² The column labeled "C4" shows the results of frame induction performed on examples from the C4 corpus, evaluated by mapping FrameNet examples to the induced clusters. The column labeled "FrameNet" shows the performance when frame induction is directly applied to. These scores are cited from the results reported by Yamada et al. (2023) and can be seen as an upper bound for our evaluation.

Overall, when fine-tuning is applied, the scores obtained using the C4 corpus are comparable to those achieved using FrameNet examples directly. This result suggests that frame induction methods based on deep metric learning are robust even when applied to real-world data. Focusing on the impact of loss functions and clustering methods, we observe that when using FrameNet examples, relatively high scores are achieved with either the Triplet or AdaCos loss in combination with twostage clustering. In contrast, when using C4 examples, the highest scores are obtained with the AdaCos loss and one-stage clustering. In addition, we observe a large performance gap between the best-performing model and the vanilla model, suggesting that deep metric learning provides a greater benefit in frame induction from real-world data.

> **Qualitative analysis** We then conducted a manual analysis of the semantic frames induced from C4 examples. We focused on the setting that achieved the highest PIF and BCF scores, using one-step clustering with the AdaCos loss. Table 3 lists examples of the induced frames along with manually assigned frame names and corresponding C4 examples.

The first two examples in Table 3 are those in which the number of associated C4 examples is approximately equal to the number of aligned FrameNet examples. For these frames, it is likely that a corresponding FrameNet frame exists. The first frame *Education_teaching* includes 'tutor' and 'train' as their frame-evoking words, and many of the corresponding FrameNet examples are annotated with the Education_teaching frame. The second frame *Violation* includes 'violate,' 'contravene,' and 'breach,' as their frame-evoking words and matches the Compliance frame, although it only covers the sense related to violation and does not include the sense related to compliance. 270

271

272

273

275

276

277

278

279

280

281

282

283

284

287

288

291

292

293

294

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

The bottom two examples in Table 3 are clusters with no aligned examples from FrameNet. These correspond to the case shown as 3b in Figure 1 and may represent novel frames not covered by FrameNet. The frame Cause_to_hasten includes 'rush,' and 'hasten' as their frame-evoking words. In FrameNet, the frames that include these verbs as LUs are limited to Self motion and Fluidic_motion, which represent voluntary actions. The causative sense of "making someone hurry," however, is not covered. The only frame-evoking verb of the frame Media streaming is 'stream.' In FrameNet, the verb stream appears as LUs only in the Mass_motion and Fluidic_motion frames, and no frame corresponding to Media_streaming is defined. The concept represented by this frame has become relatively common only in recent years, and can be regarded as a novel frame induced from real-world corpora, including recent texts.

5 Conclusion

In this study, we conducted frame induction from a real-world corpus, specifically, the Colossal Clean Crawled Corpus (C4), and performed both quantitative and qualitative evaluations by comparing the induced results with examples from FrameNet. The experimental results suggest that existing frame induction methods perform robustly even on realworld corpora. Furthermore, we found that novel frames corresponding to concepts not covered by FrameNet can also be induced. These findings indicate the potential of automatically constructing semantic frame resources for domain-specific or low-resource languages in the future.

262

263

264

265

267

²More detailed results are provided in Appendix A.

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

363

364

311 Limitations

Our study has several limitations. First, to ensure 312 that evaluation using FrameNet could be carried 313 out appropriately, we imposed a constraint such 314 that the distribution of verbs in the C4 examples used for frame induction matched the verb distribution in the FrameNet evaluation set. In real-world 317 applications of frame induction, such constraints would not be applied, and thus the results may differ slightly from those observed in our controlled experimental setup. Second, our experiments were 321 conducted exclusively on English data. It remains unclear whether the proposed approach would per-323 form similarly on other languages.

References

326

327

328

329

331

333

334

335

336

337

338

341

342

343

346

347

353

354

355

357

359

361

- Saba Anwar, Dmitry Ustalov, Nikolay Arefyev, Simone Paolo Ponzetto, Chris Biemann, and Alexander Panchenko. 2019. HHMM at SemEval-2019 task 2: Unsupervised frame induction using contextualized word embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval* 2019), pages 125–129.
- Nikolay Arefyev, Boris Sheludko, Adis Davletov, Dmitry Kharchev, Alex Nevidomsky, and Alexander Panchenko. 2019. Neural GRANNy at SemEval-2019 task 2: A combined approach for better modeling of semantic relationships in semantic frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 31–38.
- Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING 1998), pages 79–85.
 - Collin F Baker, Charles J Fillmore, and John B Lowe. 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 (ACL-COLING 1998), pages 86–90.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), pages 4171–4186.
- Charles J Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company.

- Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066.
- Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), pages 212– 220.
- Jiří Materna. 2012. Lda-frames: An unsupervised approach to generating semantic frames. In *Computational Linguistics and Intelligent Text Processing:* 13th International Conference (CICLing 2012), pages 376–387.
- Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Proceedings of the 16th European Conference on Computer Vision (ECCV 2020)*, pages 681–699.
- Dan Pelleg and Andrew Moore. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning (ICML* 2000), pages 727–734.
- Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019), pages 16–30.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Eugénio Ribeiro, Vânia Mendonça, Ricardo Ribeiro, David Martins de Matos, Alberto Sardinha, Ana Lúcia Santos, and Luísa Coheur. 2019. L2F/INESC-ID at SemEval-2019 task 2: Unsupervised lexical semantic frame induction using contextualized word representations. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval* 2019), pages 130–136.
- Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice*. International Computer Science Institute.
- Dmitry Ustalov, Alexander Panchenko, Andrey Kutuzov, Chris Biemann, and Simone Paolo Ponzetto. 2018. Unsupervised semantic frame induction using triclustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 55–62.
- Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2).

Clustering	Model	α	Pu / iPu / PiF	BCP/BCR/BCF	PIF / BCF (Yamada et al., 2023)
One-step	Vanilla	0.00	51.4 / 48.3 / 49.7	38.1 / 34.6 / 36.3	54.9 / 42.6
	Triplet	0.17	71.4 / 70.4 / 70.9	61.5 / 60.1 / 60.8	73.3 / 63.9
	Softmax	0.37	66.3 / 77.4 / 71.4	53.9 / 67.5 / 60.0	71.0 / 59.9
	AdaCos	0.37	70.0 / 76.8 / 73.3	58.7 / 67.0 / 62.6	73.5 / 62.9
Two-step	Vanilla	0.67	32.1 / 42.3 / 36.5	17.7 / 25.6 / 20.9	66.9 / 56.5
	Triplet	0.57	61.5 / 71.4 / 66.0	48.7 / 60.0 / 53.6	74.8 / 66.0
	Softmax	0.50	72.4 / 68.4 / 70.2	61.8 / 56.5 / 58.9	71.9 / 61.4
	AdaCos	0.50	71.3 / 70.2 / 70.7	60.2 / 59.3 / 59.6	75.6 / 66.2

Table 4: Detailed evaluation results of frame induction with C4 examples.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021. Semantic frame induction using masked word embeddings and two-step clustering. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), pages 811–816.

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436 437

438

439

440

- Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2023. Semantic frame induction with deep metric learning. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), pages 1833–1845.
- Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. 2019. AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR 2019), pages 10823–10832.
- Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical report, Retrieved from the University of Minnesota Digital Conservancy.

A Detailed Experimental Results

Table 4 provides the detailed results of evalua-
tion scores for our semantic frame induction ex-
periments using C4 examples. In addition to the
PIF and BCF metrics reported in Table 2, we also
present the weight α in Equation (1) and the com-
ponent scores: Purity (PU), Inverse Purity (IPU), B-
cubed Precision (BCP), and B-cubed Recall (BCR).442
443
445