Frame-Based Annotation of Multimodal Corpora: Tracking Asynchronies in Meaning Construction

XXX1[0000-1111-2222-3333]

1 XXX
xxx@xxx.xxx

Abstract. The multimodal aspects of human communication have been key in several applications of Natural Language Processing, such as Machine Translation and Natural Language Generation. Despite the recent advances in tackling multimodality in a Computational Linguistics setting, integration between NLP and Computer Vision techniques is still timid, especially when it comes to providing fine-grained accounts for meaning construction in multimodal corpora. This paper reports on an ongoing research project which aims to investigate, develop and test an appropriate methodology to annotate multimodal corpora according to a principled structured semantic representation of events, relations and entities: FrameNet.

Keywords: Frame Semantics, Multimodal Annotation, FrameNet.

1 Computational Processing of Multimodal Communication

Multimodal analyses have been growing in importance within several approaches to both Cognitive Linguistics and Natural Language Understanding, changing the scenario depicted by McKevitt (2003), according to whom little progress had been made in integrating the areas of Natural Language Processing (NLP) and Vision Processing (VP), although there had been much success in developing theories, models and systems in each of these areas separately. Turner (2018) explains that multimodality is traditionally expressed in three different forms of communication and meaning construction: auditory, visual and text. Steen et al. (2018) highlight that multimodal corpora have been annotated for correlations involving mainly gesture communication and text data, and that computational infrastructure for dealing with large multimodal corpora has been under development. In the same direction, Cohn (2016) points out that human communication is naturally multimodal and that there are significant efforts to examine semantic correspondences both in the relations between speech and gesture, as in those established between text and image. The systematization for semantic investigation proposed by this author explores the grammaticality of both modalities.

Based on Jackendoff's (2002) parallel architecture for language, Cohn (2016) focuses on how grammar and meaning coalesce in multimodal interactions, extending beyond the semantic taxonomies typically discussed about text–image relations. He thus classifies the relations between text and image in visual narratives, evaluating the

presence or absence of a grammar in the structuring of each of the modalities and also the presence or absence of semantic dominance from one of the modalities.

Although Cohn's (2016) model offers a coherent framework to approach multimodal data, the author does not incorporate any sort of fine-grained semantics into his model. Nonetheless, he recognizes the importance of using one for adequately tackling the interrelations and interactions between modalities and its components.

Given the lack of research incorporating fine-grained models of semantic cognition into multimodal analyses, the ongoing research presented in this paper aims to tackle the issue of meaning construction in multimodal settings, specifically on what concerns the interaction between audio (verbal expression transcribed into text) and video (not necessarily gesture communication), based on a principled structured model of human semantic cognition: FrameNet. Such a model is presented next.

2 FrameNet and Frame-Based Semantic Representation

Frames have a long history in both AI (Minsky 1975) and linguistics (Fillmore 1982) as structured representations of interrelated concepts. In Frame Semantics, words are understood relative to the broader conceptual scenes they evoke (Fillmore 1977). Hence, the expression child-safe beach, for example, is understood only in the context of a scene in which an Asset (the child) is exposed to some potentially Harmful_event (a strong sea current, for example). This theoretical insight is the basis for lexicographic resources such as Berkeley FrameNet and its sister projects in other languages. Currently, there are framenets for several languages, including Brazilian Portuguese[1]. These frame-based resources have been applied to different NLU problems, such as conversational AI (Vanzo et al. 2019) and paraphrase generation (Callison-Burch and Van Durme 2018).

All framenets are composed of frames and their associated roles in a network of typed relations such as inheritance, perspective and subframe. The Risk_scenario frame alluded to above, for example, is an umbrella frame for several more specific perspectivized frames such as Being_at_risk (in which the Asset is exposed to a risky situation) and Run_risk (in which a Protagonist puts an Asset at risk voluntarily). Each perspective may be evoked by different words or by one same lexeme with different syntactic instantiation patterns. Being_at_risk, for example, is evoked by adjectives such as unsafe.a and nouns such as risk.n in constructions like X is at risk. On the other hand, Run_risk is evoked by verbs such as risk.v and also by risk.n, but in a different construction: Y has put X at risk (Fillmore and Atkins 1992). The database structure also features annotated sentences, which attest the use of a given word in the target frame.

Thus, the hypothesis being investigated in this work is that, similarly to the way in which words in a sentence evoke frames and organize their elements in the syntactic locality accompanying them, video scenes may also either (i) evoke frames and organize their elements on the screen, or (ii) complement the frame evocation patterns

1 See http://globalframenet.org.

of the sentences they are attached to, providing different profiling and perspective options for meaning construction. To test the validity of this hypothesis and, therefore, the potential relevance of the project, an exploratory corpus study was conducted and is described in the next section.

3 Exploratory corpus study and annotation tool

FrameNet Brasil has been developing resources and applications for Tourism (Diniz da Costa et al. 2018). Therefore, given the existence of a fine-grained semantic infrastructure already developed for this domain, the corpus used is a Brazilian television travel show called "Pedro pelo Mundo", in which the host explores a place, highlighting its cultural and socioeconomic aspects. In the study, we annotated the audio transcript[2] using the FN-Br WebTool (Matos and Torrent, 2018), an open source database management and annotation tool that allows for the creation of frames and relations between them, as well as for the annotation of sentences and full texts. An example of the sort of text annotation carried out in this project is shown in Figure 1.

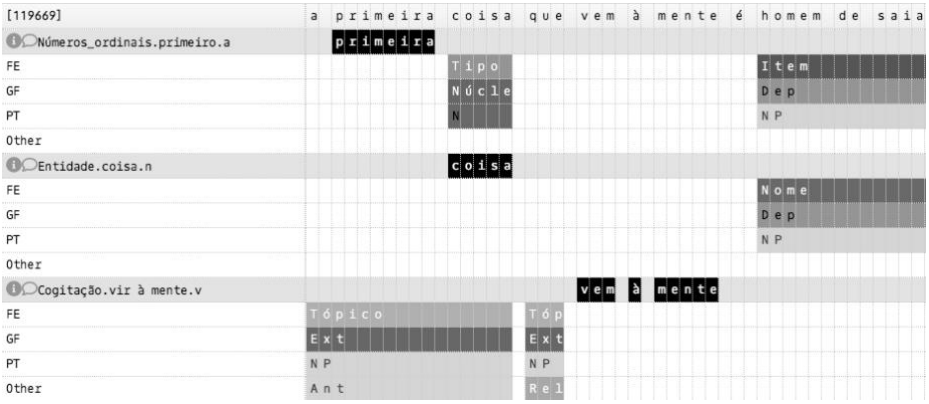

Fig. 1. Example of a sentence annotated for frames in the FN-Br WebTool

After the text/audio annotation, we annotated the video superimposed in the episodes for the same categories. Next, we contrasted the annotations, searching for matching frames while also considering the synchronicity, asynchronicity or associated synchronicity of the frames instantiated in both. The time code was taken as the correlational unit between then, as shown in Table 1.

To annotate the video, we developed a Multimodal Annotation Module for the WebTool, which allows annotators to choose frames and locate frame elements both in the text and in the images, while keeping track of the time span in which those elements are active in the video and in the audio.

2 For the examples discussed in this paper we refer to one line in the corpus: Quando a gente pensa na Escócia, **a primeira coisa que vem à mente é homem de saia, whisky escocês e gaita de fole**. *'When we think of Scotland, the first things that comes to mind are man in skirt, Scottish whisky and bagpipe'*.

Table 1. Audio (text) and Video annotation comparison.

LEXICAL UNIT	AUDIO FRAME	AUDIO TIME	VIDEO FRAME	VIDEO TIME	SYNC
primeira *(first)*	Ordinal_numbers	33.20 to 34.02	-	-	async
coisa *(thing)*	Entity	34.03 to 34.11	-	-	async
vem à mente *(come to mind*)	Cogitation	34.14 to 35.02	-	-	async
homem (*man)*	People	35.03 to 35.14	People	36.12 to 37.12	async
saia *(skirt)*	Clothing	35.17 to 35.29	Clothing	36.12 to 37.12	async
whisky	Food	36.00 to 36.10	Food	35.02 to 36.12	sync
escocês *(Scottish)*	Origin	36.11 to 36.23	-	-	async
gaita de fole *(bagpipe)*	Noise_makers	36.24 to 37.23	Noise_makers	36.12 to 37.12	sync

The preliminary results suggest that adding a multimodal domain to the linguistic layer of annotation and analysis contributes to enrich the kind of information that can be tagged in a corpus. For instance, in Figure 2, bounding box 3 is annotated for the People_by_origin frame. Such an annotation is only possible because we combine the visual data (a man wearing kilt and playing bagpipe) with the auditory data that mentions Scotland and Scottish.

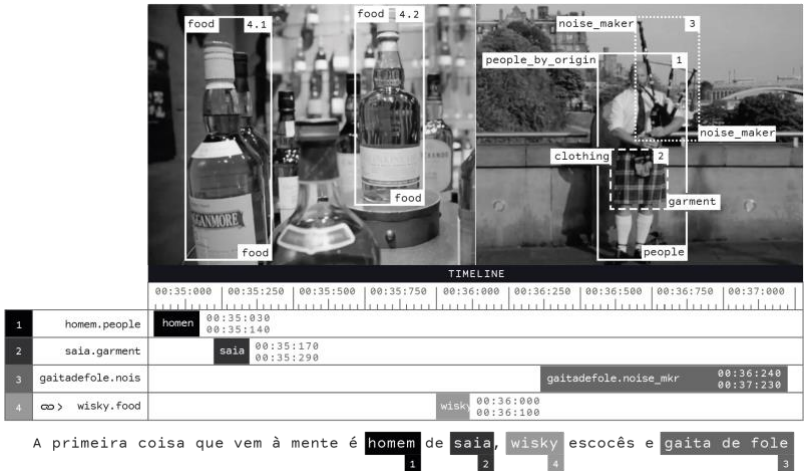

Fig. 2. Projected screen of the FN-Br Webtool Video Annotator Module.

4 Conclusions

In this paper, we presented a tool and annotation scheme for fine-grained annotation of multimodal corpora. Such a tool controls for the synchronicity between different media types and allows for cross-annotation, yielding, as an annotation product, material that can shed light on the role of multimodality in language comprehension. Future work includes the creation of a gold standard multimodal annotated corpus that may be used in Machine Learning applications such as Automatic Visual Semantic Role Labeling and video indexing.

References

1. Callison-Burch, C., & Van Durme, B. Large-Scale Paraphrasing for Natural Language Understanding. Johns Hopkins University, Baltimore (2018).
2. Cohn, N. A multimodal parallel architecture: A cognitive framework for multimodal interactions. Cognition 146, 304-323 (2016).
3. Diniz da Costa, A., Gamonal, M. A., Paiva, V. M. R. L. et al.: FrameNet-Based Modeling of the Domains of Tourism and Sports for the Development of a Personal Travel Assistant Application. In: Proceedings of the LREC 2018 Workshop International FrameNet Workshop 2018: Multilingual Framenets and Constructicons, pp 6-12. European Language Resources Association, Paris (2018).
4. Fillmore, C. J. The case for case reopened. Syntax and semantics 8, 59-82 (1977).
5. Fillmore, C. J. Frame semantics. In: The Linguistic Society of Korea (Org.). Linguistics in the Morning Calm, pp. 111-137. Hanshin Publishing Co., Seoul (1982).
6. Fillmore, C. J., & Atkins, B. T. Toward a frame-based lexicon: The semantics of RISK and its neighbors. Frames, fields, and contrasts: New essays in semantic and lexical organization 103, 75-102 (1992).
7. Jackendoff, R. Foundations of language: Brain, meaning, grammar, evolution. Oxford University Press, Oxford (2002).
8. Matos, E. E. S., Torrent, T. T. FN-Br WebTool: FrameNet Brasil Web Annotation Tool. INPI Registration Number BR512018051603-3
9. McKevitt, P. MultiModal semantic representation. In First Working Meeting of the SIGSEM Working Group on the Representation of MultiModal Semantic Information (pp. 1-16) (2003).
10. Minsky, M. A framework for representing knowledge. In P. Winston (Ed.), The psychology of computer vision (pp. 211–277). McGraw-Hill, New York (1975).
11. Steen, F. F.; Hougaard, A.; Joo, J.; et al. Toward an infrastructure for data-driven multimodal communication research. Linguistics Vanguard 4(1), - (2018).
12. Turner, M. The Role of Creativity in Multimodal Construction Grammar. Zeitschrift für Anglistik und Amerikanistik, 66(3), 357-370 (2018).
13. Vanzo, A., Bastianelli, E., & Lemon, O. Hierarchical multi-task natural language understanding for cross-domain conversational ai: HERMIT NLU. arXiv preprint arXiv:1910.00912. (2019)