

Charon: a FrameNet Annotation Tool for Multimodal Corpora

Frederico Belcavello¹, Marcelo Viridiano¹, Ely Edison Matos¹, Tiago Timponi Torrent^{1,2}

¹ FrameNet Brasil Lab, Graduate Program in Linguistics, Federal University of Juiz de Fora

² Brazilian National Council for Scientific and Technological Development – CNPq
{fred.belcavello, ely.matos, tiago.torrent}@ufjf.br, barros.marcelo@estudante.ufjf.br

Abstract

This paper presents Charon, a web tool for annotating multimodal corpora with FrameNet categories. Annotation can be made for corpora containing both static images and video sequences paired – or not – with text sequences. The pipeline features, besides the annotation interface, corpus import and pre-processing tools.

Keywords: FrameNet, Multimodality, Picture Annotation, Video Annotation, Text Annotation

1. Introduction

Multimodality refers to the property of any communication phenomenon where two or more modes – defined as experientially recognized resources for meaning-making shaped by society and culture – are brought into play (Jewitt and Kress, 2003; Kress, 2010; Bateman et al., 2017). This paper approaches the expansion of FrameNet annotation into the multimodal domain, as proposed in Belcavello et al. (2020), by presenting Charon: a semi-automatic, human-in-the-loop tool for annotating static and dynamic images for semantic frames. Charon was developed to meet the following key requirements: (i) compatibility with existing FrameNet software; (ii) annotation of image with FrameNet categories; (iii) linkage of image and textual annotations.

2. FrameNet Annotation

FrameNet is a curated language model where lexical items have their meaning defined against systems of concepts called frames (Fillmore and Baker, 2009). For instance, words such as *arrive.v* and *arrival.n* have their meanings defined based on a scene where a THEME arrives at a GOAL: the Arriving frame (Figure 1). Moreover, frames are connected to one another via a network of typed relations. The Arriving frame, for instance, is inherited by the Vehicle.landing and preceded by the Departing frames.

Annotation plays a key role in FrameNet, to the extent that it provides evidence supporting the analysis in the model. Two text annotation methods are used: lexicographic and full-text. In the former, the focus lies on a specific Lexical Unit (LU), and sentences instantiating that LU are extracted from corpora and annotated for a given frame. The aim is to cover the valence patterns of the LU, i.e. its semantic and syntactic affordances. In the latter, the focus is on the corpus being annotated, and the annotator creates Annotation Sets (AS) for each word for which there is an LU in FrameNet. Figure 2 shows two of the ASs created for the sentence in (1).

Arriving

Definition	
An object Theme moves in the direction of a Goal . The Goal may be expressed or it may be understood from context, but its is always implied by the verb itself.	
Example(s)	
Core Frame Elements	
FE Core:	
Goal [Goal] semantic_type: @location	The Goal is any expression that tells where the Theme ends up, or would end up, as a result of the motion.
Theme [Theme] semantic_type: @physical_object	The Theme is the object that moves. It may be an entity that moves under its own power, but it need not be.
Non-Core Frame Elements	
Relations	
Lexical Units	
arrival.n arrive.v come.v crest.v descend (on).v disemba	

Figure 1: The Arriving frame.

- (1) Então, acabei de chegar em Reykjavik, na Islândia.
So, I have just arrived in Reykjavik, Iceland.

In (1) the word forms *acabei* and *chegar*, highlighted in black in Figure 2, are the annotation targets. Note that, for each of them, there are three layers of annotation: Frame Element (FE), Grammatical Function (GF) and Phrase Type (PT). The column NI is used for indicating that core FEs are not instantiated in the sentence, but can be inferred.

The idea behind the development of Charon is that other communication modes, namely visual objects, can either evoke frames – similarly to LUs – or complement the valencies of LUs present in text accompanying the images (Belcavello et al., 2020), expanding FrameNet annotation to the multimodal domain. In section 4, we describe the tool, but, first, let us turn to a brief summary of other multimodal annotation tools.

[165593]	NI	Então, acabei de chegar em Reykjavik, na Islândia.
Activity_finish.acabar.v		acabei
FE	CNI	Activity
GF		Dep
PT		PSinf
Arriving.chegar.v		chegar
FE	CNI	Goal
GF		Dep
PT		PP

Figure 2: Full-text annotation for sentence (1).

3. Related Work

The past two decades have witnessed accelerated development of data labeling tools for human annotation of monomodal visual corpora – e.g. COCO Annotator (Lin et al., 2014), ImageTagger (Fiedler et al., 2018), and LabelBox (Sharma et al., 2019). Moreover, highly generic and flexible multimodal annotation tools, such as Anvil (Kipp, 2001) and ELAN (Wittenburg et al., 2006), allow users to design their own annotation schemes for timeline-based annotation of both audio and visual phenomena from multiple synchronized streams. Finally, frameworks, like SIDGrid (Levow et al., 2007), extend the functionality of ELAN by allowing the application of user-defined analysis programs to media, time series, and annotations associated with each project.

Nonetheless, none of these tools and annotation clients allows for the combination of data labeling with the extensive semantic granularity offered by the network of frames and frame elements provided by FrameNet. Allowing for such a combination is the main contribution of Charon, which is presented next.

4. Charon: Multimodal Annotation Tool

Charon is a multimodal annotation and database management tool. It was developed to annotate visual objects, correlate them with textual data and label frames and Frame Elements evoked by them. Charon is compatible with the FN-Br WebTool: a database management and annotation software used by both local framenet projects in Brazil, Sweden, Croatia and Japan, and in the Global FrameNet Shared Annotation Task (Torrent et al., 2018).¹ Charon is composed of two modules: a static mode, for annotating picture-text pairings, and a dynamic mode, for annotating video. Both are described next.

4.1. Annotation of Picture-Caption Pairings

Charon’s static annotation mode can be used to improve multimodal datasets containing picture-text pairings by adding fine-grained semantic information provided by FrameNet. The version of the tool presented in this paper has been tuned to the requirements of the

Flickr 30K Entities dataset (Plummer et al., 2015) – an expansion of Flickr 30K (Young et al., 2014) that adds manually annotated bounding boxes and coreference chains linking entities from each image to their correspondent descriptors in each caption. However, any dataset featuring pictures, captions and bounding boxes identifying parts of the picture can be used. The annotation process is divided into two stages: (i) corpus import and pre-processing, and (ii) annotation.

4.1.1. Picture Corpus Import and Pre-Processing

To upload a new corpus, all related files – a folder with JPEG images, a text file with all the sentences, and a XML with the classes and coordinates for each object’s bounding box – must be compressed into a ZIP file. Next, Charon creates a new corpus folder in which documents containing lists of image-sentence pairs are built. Before being presented to the annotator, the sentences in these documents are pre-processed by a disambiguation algorithm – DAISY (Torrent et al., 2022) – that identifies and associates each frame-evoking lemma with a semantic frame in the FrameNet database, resulting in an automated frame annotation for each sentence. Such an automated annotation can be checked during a human-in-the-loop process. Charon also checks the image related files for all objects that might have been previously tagged via data labeling tools or computer vision algorithms, and automatically correlates the classes of these objects – obtained from datasets like COCO (Lin et al., 2014) and Open Images (Kuznetsova et al., 2020) – with existing Lexical Units in FrameNet. After that, images and sentences are loaded into the interface where the human annotation happens.

4.1.2. Picture Annotation Process

Figure 3 presents the static mode interface, used for the annotation of Picture-Caption pairings. This annotation interface is composed of several panels that are loaded depending on the type of corpus or annotation task being developed. The upper left corner of the interface offers a view of the uploaded image. The panel titled Boxes shows the coordinates for the bounding boxes related to each object/entity being annotated in that picture. The Annotations panel shows the correlations between each object/entity in the image, its co-referenced phrase extracted from the sentence in the

¹The FN-Br WebTool is available at <https://github.com/FrameNetBrasil/webtool>.

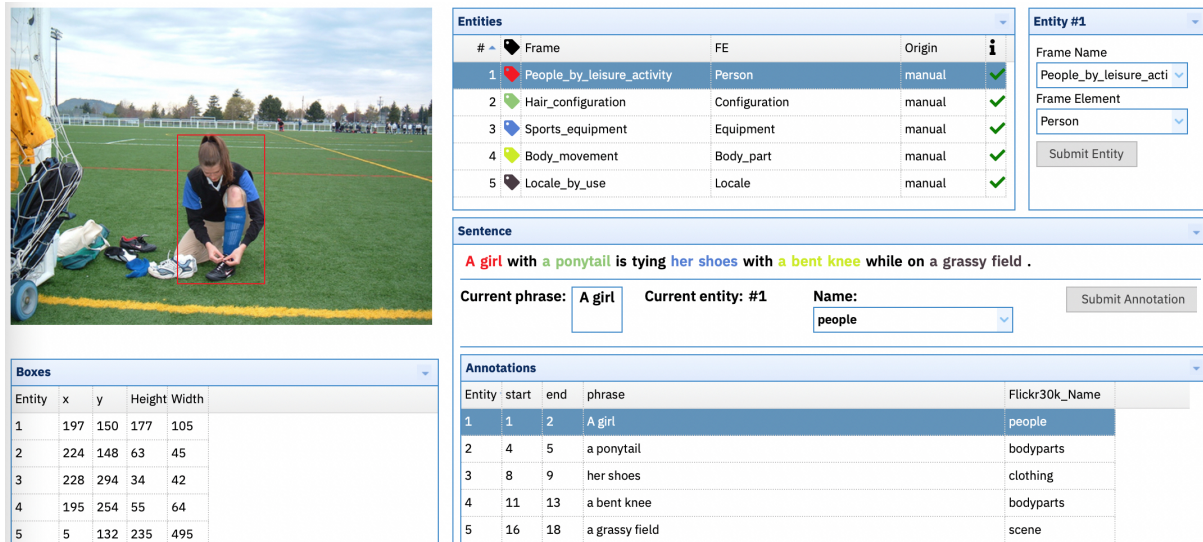


Figure 3: User interface for the annotation of Picture-Caption pairings.

middle panel, and the class used to label this object in the original dataset. Finally, the panels Entities and Object are the ones used by the human annotator to assign a Semantic Frame and a FE to each picture-text pair composed by the object/entity in the bounding box and the highlighted phrase in the sentence.

Example sentence (2) has the phrases “A girl”, “a ponytail”, “her shoes”, “a bent knee”, and “a grassy field” correlated with five distinct objects/entities in the image. For the phrase “A girl”, corresponding to the Entity 1 in the image, the annotator assigned the frame `People_by_leisure_activity` and the FE `PERSON`.

- (2) A girl in a ponytail is tying her shoes with a bent knee while on a grassy field.

This annotation mode generates an XML file, allowing the output to be used with other multimodal annotation tools and integrated with existing transcriptions and annotations from other modules and databases in the FN-Br Webtool environment.

4.2. Annotation of Videos

Two types of media are involved in the annotation of videos in FrameNet: audio and image. Therefore, this module of Charon was designed to pre-process videos by (i) extracting verbal data from both audio and images (i. e. subtitles) and deliver it for annotation in the FN-Br Webtool; and (ii) submitting image to an external computer vision system that identifies visual objects and make bounding boxes for those objects available for annotation in Charon. In the following subsections we describe the video annotation pipeline.

4.2.1. Video Corpus Import and Pre-Processing

The pipeline designed for corpus import and video pre-processing starts with the selection of the video input,

which is imported, pre-processed and separated into two data flows: one for the audio and another for the images.

The next step is the selection of the language of the verbal mode. After the language is selected, the audio data runs through a speech-to-text cloud service, which detects word by word what is said throughout the video.² Each word receives time stamps indicating the time span during which they are spoken.

From the image flow, subtitles are extracted using an optical character recognition software. They are time-stamped and then merged to the text corpus with the output of the speech-to-text software.³ Words and sentences extracted then go through a human-in-the-loop stage, where users can build sentences from the words, edit them, as well as check and adjust time stamps. Finally, the textual part of the corpus is saved and sent to the FN-Br Webtool for annotation.

Charon also processes non-verbal visual data. The images extracted at a 25 frames per second rate are stamped for both time (in seconds) and video frame (in sequential numbers). They run through a computer vision algorithm, which automatically tags objects in each frame, associating a bounding box and a category to them.⁴

At the end of the pipeline, annotators access the video annotation module, where they visualize both the annotated sentences and the automatically detected objects. This module is described next.

²For the current implementation, Google Cloud Speech API (<https://cloud.google.com/speech-to-text>) is used.

³For the current implementation, Tesseract OCR (<https://github.com/tesseract-ocr/tesseract>) is used.

⁴For the current implementation, YOLOv3 (Redmon and Farhadi, 2018), trained on the COCO dataset (Lin et al., 2014) is used.

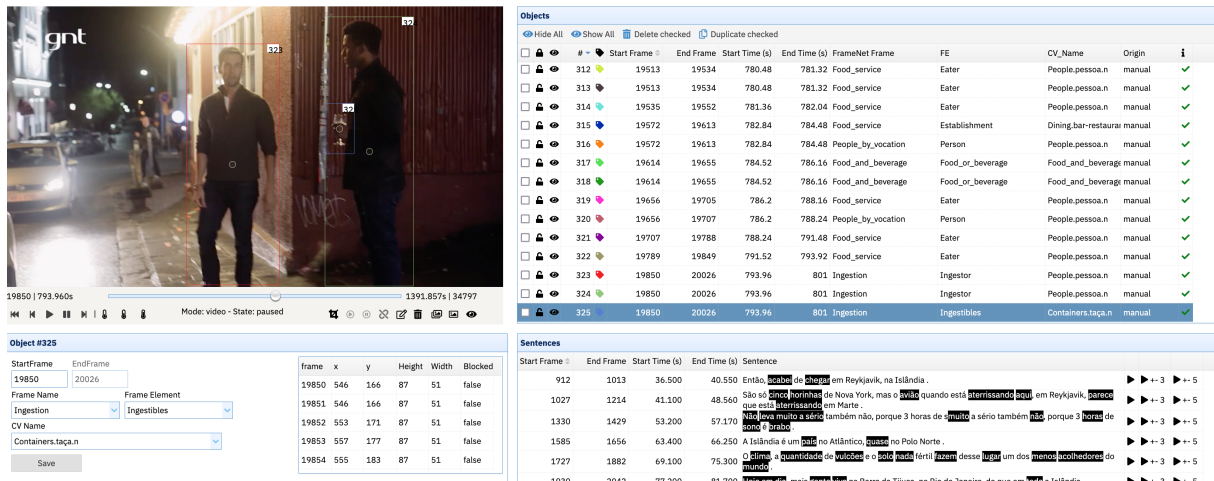


Figure 4: Example of video annotation.

4.2.2. Video Annotation Process

Charon provides a myriad of possibilities for video annotation by human users, in terms of both methodologies and goals. So far, it has been used to annotate and compare semantic frames evoked by visual objects with those evoked by LUs in sentences. This is why the video annotation module features not only the annotation tools for tagging images, but also the visualization of the sentences annotated in the FN-Br WebTool for the same corpus.

Human annotators can start by reviewing the objects automatically detected by the computer vision software. If annotators agree with the bounding box drawn by the CV software, they select the object in the panel, then use the edit tracking button in the player to link the bounding box to the object through the following video frames. Once the object is not visible anymore or there is a cut point, the annotator presses the pause tracking button, and then the end object one. If annotators do not agree with the bounding box drawn, they can select the object in the panel and delete it.

To create new objects, annotators use the new object button, draw the bounding box over the object they want to detect, then start tracking it. Tracking can be executed manually, frame by frame, or automatically, using the start tracking button. In both cases, annotators determine the end point for the bounding box when the object is not visible anymore or there is a cut point. Next, annotators have to manually attribute a Semantic Frame and a FE to the object. They choose the frame from the list under the Frame Name field. Once the frame is chosen, a list of its FEs is loaded in the Frame Element field. Annotators should also attribute a Computer Vision name to the object or confirm the label automatically assigned by the computer vision software. This category associates one LU with the object, considering its value as an entity recognizable by computer vision tools or algorithms. In the CV Name field, users may choose from any LU in the framenet database they are using. Figure 4 shows an example of video anno-

tation. At the moment the image in Figure 4 is seen on screen, viewers listen one of the men speaking the sentence annotated as in (3):

- (3) Bom^{Desirability} que aqui^{Locative_relation}
a gente bebe^{Ingestion} e vai
esquentando^{Change_of_temperature}, né?

It's good that here we drink and warm ourselves up, innit?

When looking for correspondences between text and image, objects 323 and 324 were annotated as the INGESTORS for the Ingestion frame (Figure 5). On the other hand, as what is visually recognizable are two human figures, the CV Names chosen were person.n in the People frame. Object 325 was annotated as the INGESTIBLES in the Ingestion frame and as glass.n in the Container frame for the CV Name. What is interesting here is that in the sentence there is no mention to the INGESTIBLES FE – it is a null instantiation, – neither to the Container Frame. Therefore, this example shows how meaning layers and granularity can be added to the FrameNet semantic representation by annotating visual data in correspondence with textual data in a corpus.

5. Expected Datasets

As demonstrated so far, the addition of other communicative modes to FrameNet annotation allows for building fine-grained semantically annotated multimodal datasets. Two datasets are being currently built by means of Charon's annotation affordances: the Framed Multi 30k and the Frame² datasets (Torrent et al., 2022).

The Framed Multi 30k Dataset will consist of an improved version of two datasets: the Multi30k dataset (Elliott et al., 2016) – a multilingual extension of the popular dataset for sentence-based image description

Ingestion

Definition
An Ingestor consumes food or drink (Ingestibles), which entails putting the Ingestibles in the mouth for delivery to the digestive system. This may include the use of an Instrument . Sentences that describe the provision of food to others are NOT included in this frame.
Example(s)
Core Frame Elements
FE Core:
Ingestibles [Ingestibles] The Ingestibles are the entities that are being consumed by the Ingestor .
Ingestor [Ingestor] The Ingestor is the person eating or drinking. semantic_type: @sentient
Non-Core Frame Elements
Relations
Lexical Units
dine.v down.v drink.v eat.v feast.v feed.v got

Figure 5: The Ingestion frame.

Flickr30k (Young et al., 2014), – and Flickr30k Entities (Plummer et al., 2015). For each of the 276,000 bounding boxes from Flickr30K Entities, our Framed Multi 30k dataset will add five new sets of Entity-Frame-Frame Element relations, 155,070 new Brazilian Portuguese descriptions, and 155,070 new English-Portuguese translated descriptions.

The Frame² dataset, in turn, is being built to provide means to analyze the interaction between the frame-based semantic representation of verbal language and that produced by the frame-based annotation of video sequences, i.e. sequences of visual frames related with audio, forming a video. The aim is to make it possible to analyze audio and video combination possibilities in terms of frames, as in the example shown in Figure 4. This dataset is composed by the multimodal objects selected for annotation in the corpus of the TV Travel Series “Pedro pelo Mundo.” The first data release of Frame² will comprise the annotation of all 10 episodes of the show’s first season. This means approximately 12,200 annotation sets for text and 5,000 for image.

6. Conclusion

Charon is a unique and robust tool that provides an user-friendly, web-based interface for fine-grained semantic annotation of both static and dynamic multimodal corpora. The integration with the ever-growing network of semantic frames provided by framenets worldwide allows for large-scale multimodal data analysis. While the current release has already demonstrated its usefulness, many updates and extensions are in the works. A priority is to improve the integration with metadata obtained from machine vision models for automatic object detection.

7. Acknowledgments

Authors acknowledge the support of the Graduate Program in Linguistics at the Federal University of Juiz de Fora, as well as the role of Oliver Czulo and Mark Turner in co-supervising, respectively, the doctoral research by Marcelo Viridiano and Frederico Belcavello, whose contributions are reported in this paper. Research presented in this paper was funded by CAPES PROBRAL grant 88887.144043/2017-00 and CNPq grants 408269/2021-9 and 315749/2021-0. Viridiano’s research was funded by CAPES PROBRAL PhD exchange grant 88887.628830/2021-00. Belcavello’s research was funded by CAPES PDSE PhD exchange grant 88881.362052/2019-01. Authors also acknowledge the contributions of E. P. Hackett⁵ and Prishita Ray⁶ to the initial development phase of Charon. Both projects were developed and funded under the Google Summer of Code Program.

8. Bibliographical References

- Bateman, J., Wildfeuer, J., and Hiippala, T. (2017). *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Mouton Textbook. De Gruyter.
- Belcavello, F., Viridiano, M., Diniz da Costa, A., Matos, E. E. d. S., and Torrent, T. T. (2020). Frame-based annotation of multimodal corpora: Tracking (a)synchronies in meaning construction. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 23–30, Marseille, France, May. European Language Resources Association.
- Elliott, D., Frank, S., Sima’an, K., and Specia, L. (2016). Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August. Association for Computational Linguistics.
- Fiedler, N., Bestmann, M., and Hendrich, N. (2018). Imagetagger: An open source online platform for collaborative image labeling. In *Robot World Cup*, pages 162–169. Springer.
- Fillmore, C. J. and Baker, C. (2009). A frames approach to semantic analysis. In Bernd Heine et al., editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK, December.
- Jewitt, C. and Kress, G. (2003). *Multimodal Literacy*. New literacies and digital epistemologies. P. Lang.
- Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*. Citeseer.

⁵<https://summerofcode.withgoogle.com/archive/2019/projects/5902293138931712>

⁶<https://summerofcode.withgoogle.com/archive/2020/projects/4857286331203584>

- Kress, G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. Multimodality: A Social Semiotic Approach to Contemporary Communication. Routledge.
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.
- Levow, G.-A., Bertenthal, B., Hereld, M., Kenny, S., McNeill, D., Papka, M., and Waxmonsky, S. (2007). Sidgrid: A framework for distributed and integrated multimodal annotation and archiving and analysis. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 231–234.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In David Fleet, et al., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.
- Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- Sharma, M., Rasmuson, D., Rieger, B., Kjelkerud, D., et al. (2019). Labelbox: The best way to create and manage training data. software, labelbox. Inc, <https://www.labelbox.com>.
- Torrent, T. T., Ellsworth, M., Baker, C., and Matos, E. E. d. S. (2018). The Multilingual FrameNet Shared Annotation Task: a Preliminary Report. In Tiago Timponi Torrent, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).
- Torrent, T. T., Matos, E. E. d. S., Belcavello, F., Viridiano, M., Gamonal, M. A., Costa, A. D. d., and Marim, M. C. (2022). Representing context in framenet: A multidimensional, multimodal approach. *Frontiers in Psychology*, 13.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).
- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.