NUANS - MiniHomework 1b: NounAtlas

Marco Lo Pinto

1 Introduction

In the field of Natural Language Processing, a linguistic inventory is used for a variety of tasks, such as Semantic Role Labeling. The choice of a particular inventory with respect to another could depend on the coverage of word meanings, the type and number of clusters used to classify, the number of semantic roles and so on. One of them, called VerbAtlas (Di Fabio et al., 2019) offers, with respect to other inventories such as FrameNet (Baker, 2014) and PropBank (Palmer et al., 2005), cross-domain roles, semantically consistent frames and informative role labels.
One main problem concerning VerbAtlas is that it's only verb-specific. In order to improve it further, the aim of this paper is to enhance the inventory with nominal events synsets, starting from the VerbAtlas' frames.

2 Linking together VerbAtlas, BabelNet and WordNet

2.1 From frames to synsets

In VerbAtlas, each frame corresponds to a cluster of BabelNet synsets and each BabelNet synset can be associated with a WordNet synset. Using these informations, a connection between these three linguistic resources can be generated. After doing that, we can observe that not all VerbAtlas frames are populated with BabelNet synsets: 425 VerbAtlas frames are mapped into BabelNet synsets out of 432.

2.2 Exploring WordNet graph

Nonetheless, a first exploration of the WordNet nominal graph is done: starting from VerbAtlas frames, each BabelNet synset is converted into the corresponding WordNet. Then, by exploiting NLTK (Loper and Bird, 2002), we can move from verbal to nominal synsets using their derivationally related forms (e.g. eat (v) $\rightarrow$ eating (n), eater (n)). In order to go up in the graph using hypernyms and be sure to have explored every crucial part of it (a synset can have multiple hypernyms), a classic Depth First Search (Wikipedia, 2022) is applied, categorizing each encountered synset by the number of times that is explored and its definition.

2.3 Selection of best hypernyms

From the obtained list of synsets, its definiton and number of visits, the next step is to identify which of the derivationally related forms (and their hypernyms) are events. To do so, by ordering the explored synsets in descending order (i.e. from the most visited to the least visited) I classified the most predominant and pertinent synsets that represent events. The steps to do so are as follows:

1. Select the most predominant synsets that represent events (e.g. event.n.01 or act.n.02).
2. Explore again the graph and remove from the output all synsets that are hyponyms of the selected synsets.
3. Do again these steps until there are no more synsets that represent events.

2.4 Selection of candidate synsets

The graph is explored again using DFS so to obtain candidate synsets that are connected with a derivationally related form present in a VerbAtlas frame. The logic behind it is that if via recursive hypernymy a synset is in the predominant synsets (obtained in the last section), then that synset is a possible candidate for a VerbAtlas frame.

2.5 Resolving ambiguity

There is one main problem that was not addressed so far: what if a synset is ambiguous? (i.e. it can be in multiple frames?). For example, the synset articulation.n.03 is connected via derivationally related forms with synsets in frames EXPLAIN, SPEAK

and PRONOUNCE (it can be seen in figure 2). Various solutions can be proposed, but the one chosen in this paper is to put the ambiguous synset in a particular frame using some heuristics. More precisely, by using statistics, the frame in which the derivationally related form is connected the most (or its hypernyms if there is uncertainty), it will be put in it.

3 Results and Conclusions

Around 3974 synsets were found and saved, with 42 VerbAtlas frames that are not connected with nominal events out of 432. The number of ambiguous candidates solved is around 1075. Some clarifications must be done in the resolution of ambiguous synsets: solutions involving ML models were excluded because a well-defined statistic approach is less prone to errors. Moreover, another (feasible) approach could be to join together some frames: this can be done starting from the nominal synsets and the frames directly associated with them. A visual representation of a VerbAtlas frame and its connection with WordNet can be seen in figure 1.

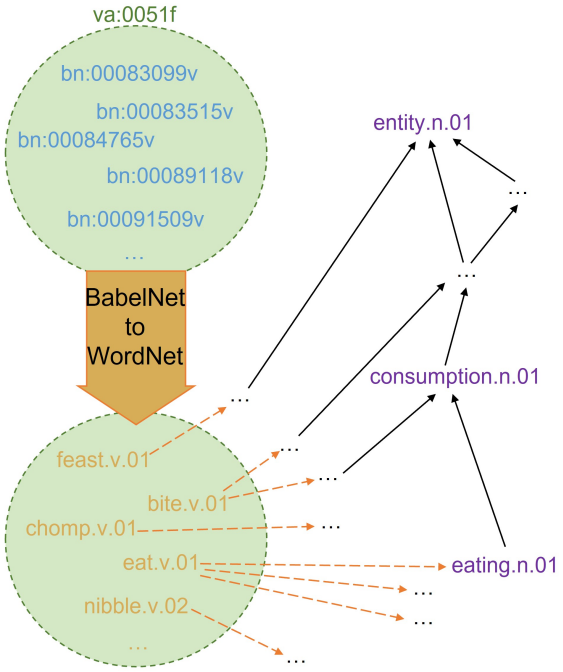

Figure 1: Visual representation of the mapping and exploring part for VerbAtlas (green), BabelNet (blue) and WordNet (brown), va:0051f is EAT-BITE VerbAtlas frame. Dotted arrows represent the semantically related forms connections, whereas uniform arrows represent hypernyms connections.

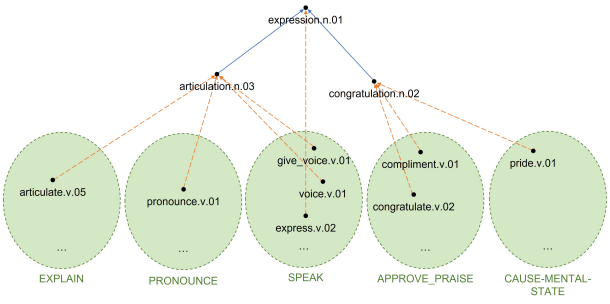

Figure 2: Visual representation of the articulation.n.03 synset ambiguity. Dotted arrows represent the derivationally related forms connections, whereas uniform arrows represent hypernyms connections.

References

Collin F. Baker. 2014. FrameNet: A knowledge base for natural language processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5, Baltimore, MD, USA. Association for Computational Linguistics.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *CoRR*, cs.CL/0205028.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Wikipedia. 2022. Depth-first search — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Depth-first%20search&oldid=1117448546. [Online; accessed 04-November-2022].