Woolery: Extending Frame Semantics to Structured Documents

Chloe Eggleston
College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst, USA
ceggleston@umass.edu

Jeremy Abramson
Information Sciences Institute
University of Southern California
Los Angeles, USA
abramson@isi.edu

***Abstract*—This paper presents Woolery, a system for semantic annotation and mapping of structured documents (such as JSON key-value pairs) to FrameNet. Implemented as a graphical interface, Woolery provides an annotator with a guided means to map keys in a JSON document to FrameNet elements, without the need for extensive knowledge of FrameNet's semantic structures. Candidate frame elements are identified via a search across FrameNet's internal representations, or via mapping keys to their potential WordNet synsets. Final element selection is automated via a pretrained language model. Initial results are promising, with the model giving an overall accuracy of 77.8% when labeling frames across a diverse corpus of JSON document schemas.**

***Index Terms*—FrameNet, natural language processing, annotation, lexical databases, JSON, ontology alignment, computational semantics**

I. Introduction

Machine-readable, structured documents such as JSON are ubiquitous. While these documents have notional semantics — for example, the implicit "is a" relationship from many JSON key/value pairs — they often lack explicit semantics, especially *linguistic* semantics. Extracting linguistic semantics from unstructured text is a fertile area of research [1] and *FrameNet* [2] is a common target. However, semantic inference on structured documents such as JSON is less studied, and poses unique challenges.

This paper presents Woolery, an annotation platform that assists a user in quickly "marking up" the keys in a JSON document with their associated FrameNet elements, surfacing the document's inherent semantics. The FrameNet mapping provided by Woolery is useful for many downstream tasks, such Semantic Role Labeling [3], machine translation [4], and rule-based Natural Language Generation systems [5].

This paper makes the following key contributions:

- Woolery, an open-source graphical annotator for JSON documents. Woolery maps the keys in these documents to their appropriate FrameNet lexical units (LUs)
- A transformer-based LU selection model, for automated annotation
- Evaluation across a robust set of differing JSON documents

This work was performed in a Research Experiences for Undergraduates program funded by National Science Foundation grant #1659886.

FrameNet is an English lexical database used in the natural language processing community to label semantic frames: recurring linguistic structures that organize concepts and events into a defined schema. It provides a theory of semantics in which the meanings of a given word are dependent on its context [6]. A common way of ascribing linguistic semantics to text is via *Frame Semantic Parsing*, a task in which frame structures are extracted from text [1]. There are three fundamental parts of frame semantic parsing: *target identification* — selecting specific words that can evoke frames, *frame identification* — choosing the specific frame(s) evoked by said words, and *argument identification* — finding out how the lexical units fit into a given frame's argument structure. Woolery focuses on the second task.

A semantic frame in FrameNet consists of *frame elements* (FE) — contextual arguments that emphasize certain words' roles in a frame — and *lexical units* (LU) — specific words that evoke a given frame, which are paired with the potential FE(s) they align to and the sources that they were annotated from. Woolery seeks to intelligently assist a human annotator in mapping the keys of a JSON document to the most appropriate LU for the context.

For example, consider the case of a key called `height`, used in the following contexts `{"height": "29032 ft"}` and `{"height": "200 px"}`. FrameNet provides two distinct lexical units for height: one from the frame `Natural_features` and one from the frame `Dimension`. The former concerns height within the *altitude* frame, and the latter within the frame describing an object's *scale*. The contextual nature of frame semantics when paired with the key/value pair's structure allows for classification of words into their meanings, providing a format for semantic annotation and an endpoint for word sense disambiguation models. Woolery leverages this context to intelligently surface the appropriate LUs and frames for a given key, and allow the user to select which is most appropriate, if necessary.

The rest of this paper is structured as follows: Related work is presented in the next section. In section III, we discuss the current implementation of Woolery, and the techiques used for intelligent mapping. Section IV outlines preliminary results, mapping a diverse corpora of JSON documents using Woolery.

We discuss potential future avenues of research and conclude in section V.

II. Related Work

In the context of FrameNet, much previous work in the annotation space focuses on frame identification, as it provides a frame-semantic analogue to word sense disambiguation, or argument identification, being a form of semantic role labeling. For example, the need to contextualize software requirements was addressed in such a manner with a semi-automated approach [7].

In addition, fully automated approaches to both rely on the advancements of their generalized field; frame identification models such as [8] adapt word sense disambiguation approaches like [9] to a frame-centric domain.

The complementary nature of these two perspectives has been noted in natural language processing discourse [10], with emphasis on how FrameNet can be augmented using WordNet [11], a lexical resource focused on word senses as opposed to semantic contexts like FrameNet.

III. Design and Implementation

The Woolery annotator is implemented as a graphical user interface (see Figures 1 and 3). Woolery recurses down the tree structure of a JSON document, building a list of keys. These keys are tokenized into words, and each of these words are stemmed. For example, the key-value pair: `{"DateCreated": 1627776000}` would be tokenized into `date` and `created` and stemmed into `date` and `creat`. The user can iterate through the keys of the document — or choose a specific one — and then choose from the different mapping approaches outlined in this section.

The first approach uses a regex search of the stemmed/tokenized key(s), across of all the LUs in FrameNet, and provides candidates that match. For the example above, the following LU candidates are obtained: *date.n*, *date.v*, *consolidate.n* (etc.) for `date`, and *create.v*, *creation.n* for `creat`. After examining the related linguistic and contextual information about the given LU and its frame (as shown in Figure 1), the user can select the appropriate annotation.

While such a search is often sufficient for extracting the semantics of a JSON key, FrameNet may not include the specific term in its lexical units. To mitigate this, Woolery can search synonyms and hypernyms (or "umbrella terms") for matches via WordNet [11]. WordNet is constructed of *synsets*, or labeled groups of synonyms. Each of these synonyms, called lemmas, share the same part of speech and lexical function, as well as links to other synsets of hypernyms that are a broader related concept. For example, "animal" is a hypernym of "horse", as animal is a more broad term that can serve the same lexical function.

Woolery feeds the terms through WordNet's Morphy interface[1] and then performs a search to create a list of potential synsets. These are split into synset categories called lexnames, such as *noun.object*, *noun.quantity*, *verb.change*, and *verb.cognition*. The user selects one of these categories and is presented with all the related synsets within this category, as well as a variety of linguistic information, allowing them to make an informed selection.

However, this selection results in a WordNet synset instead of a FrameNet lexical unit. In order to address this, we utilized mappings from WordNet to FrameNet. The majority of our WordNet to FrameNet mappings were provided from the Framester project [12]. We aggregated these mappings using Framester's curated data, as well as data from Framebase [13], and the data from Extended WordFrameNet (XWFN) [14]. While the latter two were provided as alternate test sets to the Framester mappings, we took the union of all three sources to provide a more comprehensive mapping. Coverage results are presented in Table I. The table shows the number of WordNet synsets we were able to map to FrameNet LUs, where the Hyp_n notation represents hypernyms of those mappings, as well as hypernyms of hypernyms of those mappings, up to n times (inclusive). Even without "compound" hypernyms, this approach surfaced almost $10,000$ potential synsets that have an LU mapping.

Source	Direct	Hyp_1	Hyp_3
Framester	8,787	28,717	55,341
Framebase	7,402	27,008	53,792
XWFN	6,261	23,592	4,8921
Union	9,939	31,387	58,778
(All Synsets)	117,659	117,659	117,659

TABLE I
WordNet ↔ FrameNet Coverage

Since Framester provides LUs from FrameNet 1.5 and Woolery uses FrameNet 1.7, we were required to construct a mapping interface to this newer format. For frames in FrameNet 1.5 not in FrameNet 1.7, we checked if there was a 1.7 frame with the same lexical units as the mentioned 1.5 frame. This, along with manual verification, allowed us to map some frames that were renamed or altered between the versions. Entries where no frame match was found were not included, and we used the word/part of speech pair to obtain the LU mappings within each given frame.

Lastly, in addition to the other guided searches, Woolery provides a directed search capability, allowing the user to both free-text search and traverse the hierarchy of all of FrameNet, and explicitly select the lexical unit that is the most desirable for the given key.

A. Automated Lexical Unit Selection Via Transfer Learning

While the aforementioned approaches can help an annotator quickly surface candidate LUs, choosing the best match can be cumbersome, especially if one is not familiar with the internals of FrameNet. To help automate this process, Woolery leverages a language model that takes a set of candidate LUs and attempts to select which one is closest to the key's intended meaning in context.

To do this, the model (see Figure 2) first pre-processes the original key. If the key is "multiple" words — such as camel

[1] https://wordnet.princeton.edu/documentation/morphy7wn

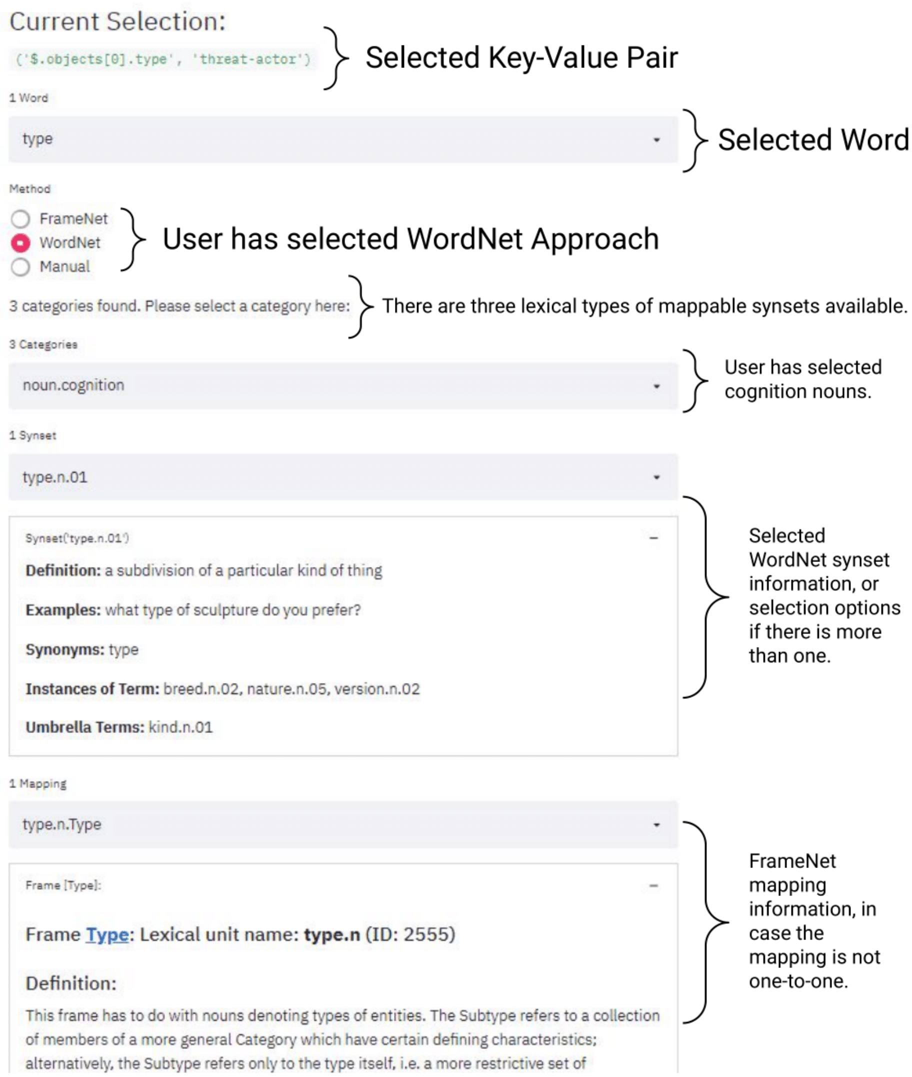

Fig. 1. Woolery WordNet Interface

case, snake case, and other forms commonly seen in JSON — it is split into multiple sub-keys and iteratively fed into the model (along with its candidate LUs). In addition, if the key is an integer, we convert it to a text representation. Next, we narrow down the possible candidates using a regex search for those starting with either the stemmed key or the key fed through WordNet's Morphy interface. Then, we feed each potential LU into a frame disambiguation transformer model based on [8] and [9].

This is implemented through inserting the key and value into a natural language prompt, then adding a `[SEP]` token followed by the LU's definition and the definition of the frame it evokes. These prompts are each tokenized and fed into BERT [15], a pretrained transformer-based language model. This is then piped into a linear layer, which returns a value for each LU candidate. The model outputs are then concatenated and fed into a softmax loss function, which is trained through backpropagation, and is capable of inference through taking the candidate with the highest probability (see Figure 3).

IV. Results and Limitations

To validate our approach, we annotated examples from the following public JSON APIs: Coin Gecko[2], D&D 5e[3], Free Dictionary[4], Metropolitan Museum of Art[5], Open Trivia[6], STIX[7], and XKCD[8] JSONs. The JSON from our evaluation suite are all from different schemas, with a wide variety of keys and "latent" semantics. We trained this using the AdamW optimizer [16] with a learning rate of 1e-5 and a 80-20 train-test split for 3 epochs.

[2]https://www.coingecko.com/en/api/documentation
[3]http://www.dnd5eapi.co/
[4]https://dictionaryapi.dev/
[5]https://metmuseum.github.io/
[6]https://opentdb.com/api_config.php
[7]https://github.com/oasis-open/cti-stix2-json-schemas
[8]https://xkcd.com/json.html

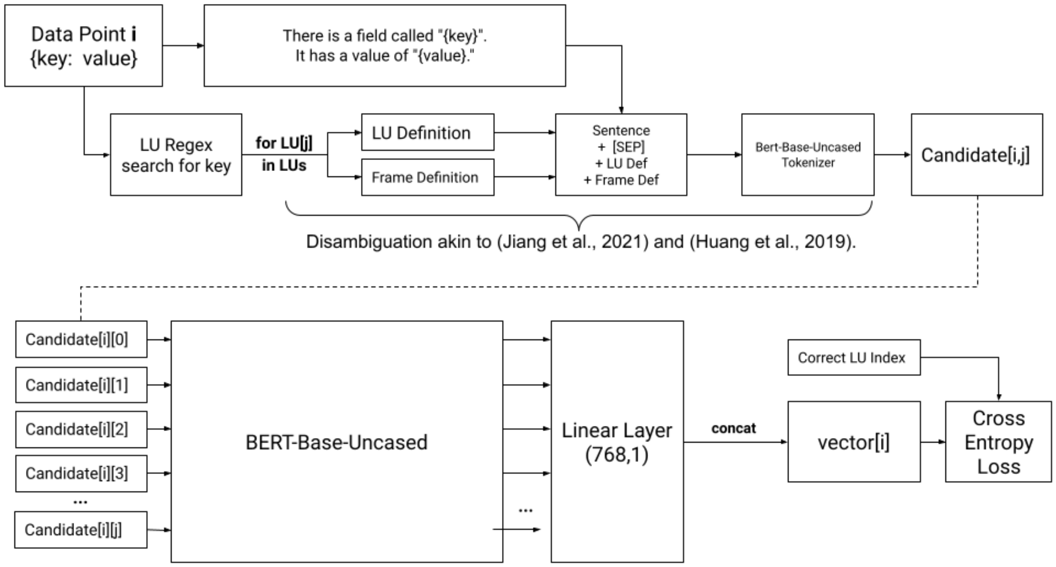

Fig. 2. Woolery Predictor Model Architecture

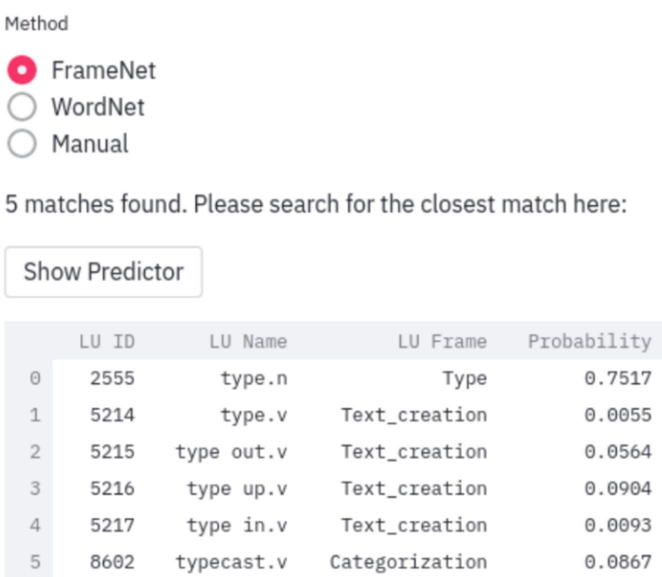

Fig. 3. Woolery Annotator Integration

Source	LU Acc.	Frame Acc.
FrameNet 1.7 (Test Split)	0.872	0.890
Annotated JSONs		
CoinGecko	0.971	0.971
D&D 5e	0.857	0.893
FreeDict	0.917	1.0
Met. Museum of Art	0.520	0.560
OpenTrivia	0.769	0.846
STIX	0.630	0.652
XKCD	0.333	0.333
Total JSONs	0.734	0.778

TABLE II
WOOLERY PREDICTOR RESULTS

As shown in Table II, initial results are promising. Notably, the FreeDict document showed perfect frame accuracy. This is likely due to the abundance of linguistics-related terms in the JSON keys, and their corresponding occurrences in FrameNet. Conversely, documents from more niche sources — such as the cyber security-related STIX, or the tech webcomic XKCD — likely have many symbols or phrases that neither FrameNet nor the transformer model could interpret. This is further illustrated with the D&D 5e and CoinGecko documents, both of which use rather domain-agnostic language, whereas the Metropolitan Museum of Art schema has many art-specific terms and distinct contextual meanings.

We have observed three independent elements that account for the model's performance: the regex search, the model's disambiguation architecture, and the prompt that the JSON has been processed in. Recent research [17] indicates that a prompt can be worth many examples worth of training. Studying the performance of additional prompts is an area for future work. In addition, there are many domain-specific keys such as "URL", "website", "email", and "screenshot" that *no* model could properly align, as they do not exist currently in FrameNet.

V. Future Work and Conclusion

While the work presented here is promising, there are a number of avenues for potential improvement. If provided with a larger corpus of data, leveraging the semantics of the JSON values themselves via (e.g.) semantic role labeling could improve mapping accuracy. In addition, the hierarchy of JSON keys has semantics, for example, nesting a `{"state"}` key under a `{"location"}` key. Leveraging the semantics of multiple JSON keys could also improve performance. Lastly, while we have tested a suite of diverse JSON schemas, we have not extensively tested numerous documents of the same type. This is an avenue we are currently exploring.

With the ubiquity of structured documents on the Internet, easily attaching Frame Semantics to their structure is of great benefit. However, determining the semantics of key/value pairs in structured documents can be a time consuming task. Mapping documents to FrameNet requires either automated approaches — which are often inaccurate or need large corpora of unstructured texts — or a user with intimate knowledge of its internals. Intelligent annotation with Woolery helps mitigate the deficiencies of fully automated annotation approaches, without requiring linguistic expertise on behalf of the user.

The issue of domain specific acronyms such as URL brings up a potential additional application of the disambiguation architecture mentioned. If one could extract a list of potential acronyms in context and use a similar definition based disambiguation, they could theoretically align algorithms to a variety of lexical resources such as FrameNet and WordNet.

Another factor to consider is how different valued documents using a given schema affect the accuracy of the classification. From our small test set, the model seems to take the value into account (as in it is less accurate without a value for the key), but most of the time, different results with the same key tend to be classified as the same lexical unit (with a few exceptions).

Provided with a corpus of frame-annotated JSONs, we hope to enable a variety of frame semantic tasks on this new domain, such as Target Identification and Argument Identification. Moreover, this domain allows for the joint utilization of graphical machine learning methods with natural language methods, possibly allowing for pretrained models that might benefit both fields.

References

[1] D. Das, D. Chen, A. F. Martins, N. Schneider, and N. A. Smith, "Frame-semantic parsing," *Computational linguistics*, vol. 40, no. 1, pp. 9–56, 2014.

[2] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. [Online]. Available: https://aclanthology.org/C98-1013

[3] M. Kshirsagar, S. Thomson, N. Schneider, J. Carbonell, N. A. Smith, and C. Dyer, "Frame-semantic role labeling with heterogeneous annotations," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 218–224. [Online]. Available: https://aclanthology.org/P15-2036

[4] S. Peron-Corrêa, A. Diniz, M. Lara, E. Matos, and T. Torrent, "Framenet-based automatic suggestion of translation equivalents," in *Computational Processing of the Portuguese Language*, J. Silva, R. Ribeiro, P. Quaresma, A. Adami, and A. Branco, Eds. Cham: Springer International Publishing, 2016, pp. 347–352.

[5] D. Dannélls and N. Gruzitis, "Controlled natural language generation from a multilingual framenet-based grammar," in *Controlled Natural Language*, B. Davis, K. Kaljurand, and T. Kuhn, Eds. Cham: Springer International Publishing, 2014, pp. 155–166.

[6] C. J. Fillmore, "Frame semantics and the nature of language," in *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, vol. 280, no. 1. New York, 1976, pp. 20–32.

[7] R. B.-n. Waad Alhoshan and L. Zhao, "A framenet-based approach for annotating software requirements." in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, T. T. Torrent, L. Borin, and C. F. Baker, Eds. Paris, France: European Language Resources Association (ELRA), may 2018.

[8] T. Jiang and E. Riloff, "Exploiting definitions for frame identification," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 2429–2434. [Online]. Available: https://aclanthology.org/2021.eacl-main.206

[9] L. Huang, C. Sun, X. Qiu, and X. Huang, "GlossBERT: BERT for word sense disambiguation with gloss knowledge," 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3509–3514. [Online]. Available: https://aclanthology.org/D19-1355

[10] C. F. Baker and C. Fellbaum, "WordNet and FrameNet as complementary resources for annotation," in *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 125–129. [Online]. Available: https://aclanthology.org/W09-3021

[11] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[12] A. Gangemi, M. Alam, L. Asprino, V. Presutti, and D. R. Recupero, "Framester: A wide coverage linguistic linked data hub," in *Knowledge Engineering and Knowledge Management*, E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, Eds. Cham: Springer International Publishing, 2016, pp. 239–254.

[13] J. Rouces, G. de Melo, and K. Hose, "FrameBase: Representing n-ary relations using semantic frames," in *Proceedings of ESWC 2015*, 2015, pp. 505–521.

[14] E. Laparra and G. Rigau, "eXtended WordFrameNet," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/799_Paper.pdf

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[17] T. Le Scao and A. Rush, "How many data points is a prompt worth?" in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 2627–2636. [Online]. Available: https://aclanthology.org/2021.naacl-main.208