The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

Extracting Academic Senses: Towards An Academic Writer's Dictionary

Hsin-Yun, Chung
NLP Lab
National Tsing Hua University
maggie100573@gmail.com

Li-Kuang Chen
Institute of
Information Systems
and Applications
National Tsing Hua University
lkchen@nlplab.cc

Jason S. Chang
Department of
Computer Science
National Tsing Hua University
jason@nlplab.cc

Abstract

We present a method for determining intended sense definitions of a given academic word in an academic keyword list. In our approach, the keyword list are converted into unigram of all possible Mandarin translations, intended or not. The method involve converting words in the keyword list into all translations using a bilingual dictionary, computing the unigram word counts of translations, and computing character counts from the word counts. At run-time, each definition (with associated translation) of the given word is scored with word and character counts, and the definition with the highest count is returned. We present a prototype system for the Academic Keyword List to generate definitions and translation for pedagogy purposes. We also experimented with clustering definition embeddings of all words and definitions, and identifying intended sense in favor of embedding in larger clusters. Preliminary evaluation shows promising performance. This endeavor is a step towards creating a full-fledged dictionary from an academic word list.

Keywords: word sense disambiguation, academic writing, academic keywords

1 Introduction

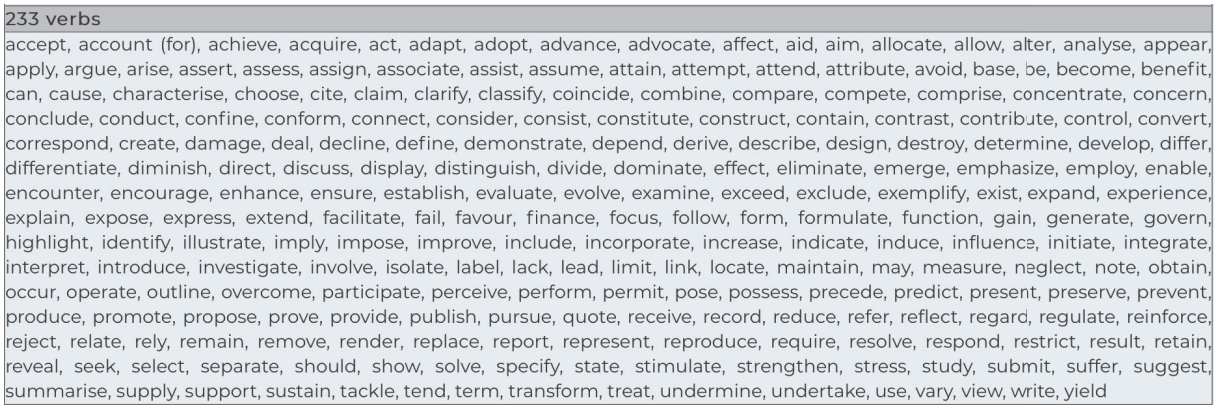
Many learners of English as a second language are writing in English for academic purposes (EAP) (e.g., papers, grant proposals, essays) every day, and many academic word lists specifically target vocabulary for EAP. For example, Academic Word List (Coxhead, 2000)[1] was developed from a small corpus (3.5 m words) of written academic English by computing the frequency, range, and uniformity of occurrence of words not in the General Service List of 2,000 words (West, 1953).

Word lists such as the AWL (Coxhead, 2000) often come without definitions (Figure 1)[2]. However, the best vocabulary learning materials for the EAP learners should contain not just words (e.g., *propose* or *argument*), but also the senses of the word (e.g., propose intended as "to offer or suggest a possible plan or action") that are relevant in an academic discourse (e.g. "We propose a method for ⋯"). Intuitively, by identifying senses that are the most similar to other words (e.g., present, introduce, and describe) in the list with an EAP prospective, we can identify the intended sense of a given word in an EAP word list. Similarity can be measured in terms of translation, definition wording, and word embeddings.

Consider the word "propose" in AKL. The best "academic" sense of this word is probably not "to ask someone to marry you", but rather "to offer or suggest a possible plan or action." The intended academic senses are typically "used to refer to those activities that characterize academic work, organize scientific discourse and build the rhetoric of academic texts" (Paquot, 2010). These senses can be identified by computing the counts of their translations. Intuitively, by requiring the senses to have translations that are shared by many words in the academic keyword list, we can bias the sense disambiguation process towards identifying these academic senses so as to best facilitate the vocabulary building and provide the learners with deep vocabulary knowledge of the academic words. Details and examples of our method will be described in

[1]The AWL can be found at www.wgtn.ac.nz/lals/resources/academicwordlist

[2]The entire AKL can be found at uclouvain.be/en/research-institutes/ilc/cecl/academic-keyword-list.html


233 verbs

accept, account (for), achieve, acquire, act, adapt, adopt, advance, advocate, affect, aid, aim, allocate, allow, alter, analyse, appear, apply, argue, arise, assert, assess, assign, associate, assist, assume, attain, attempt, attend, attribute, avoid, base, be, become, benefit, can, cause, characterise, choose, cite, claim, clarify, classify, coincide, combine, compare, compete, comprise, concentrate, concern, conclude, conduct, confine, conform, connect, consider, consist, constitute, construct, contain, contrast, contribute, control, convert, correspond, create, damage, deal, decline, define, demonstrate, depend, derive, describe, design, destroy, determine, develop, differ, differentiate, diminish, direct, discuss, display, distinguish, divide, dominate, effect, eliminate, emerge, emphasize, employ, enable, encounter, encourage, enhance, ensure, establish, evaluate, evolve, examine, exceed, exclude, exemplify, exist, expand, experience, explain, expose, express, extend, facilitate, fail, favour, finance, focus, follow, form, formulate, function, gain, generate, govern, highlight, identify, illustrate, imply, impose, improve, include, incorporate, increase, indicate, induce, influence, initiate, integrate, interpret, introduce, investigate, involve, isolate, label, lack, lead, limit, link, locate, maintain, may, measure, neglect, note, obtain, occur, operate, outline, overcome, participate, perceive, perform, permit, pose, possess, precede, predict, present, preserve, prevent, produce, promote, propose, prove, provide, publish, pursue, quote, receive, record, reduce, refer, reflect, regard, regulate, reinforce, reject, relate, rely, remain, remove, render, replace, report, represent, reproduce, require, resolve, respond, restrict, result, retain, reveal, seek, select, separate, should, show, solve, specify, state, stimulate, strengthen, stress, study, submit, suffer, suggest, summarise, supply, support, sustain, tackle, tend, term, transform, treat, undermine, undertake, use, vary, view, write, yield

Figure 1: A sample of the AKL.
From https://uclouvain.be/en/research-institutes/ilc/cecl/academic-keyword-list.html

Section 3.

The rest of the paper is organized as follows. We review the related work in the next section. Then we present our method for automatically learning to determine intended senses of words of a keyword list for a specific genre (Section 3). As part of our evaluation, we compare the quality of the senses generated by the proposed method with a clustering based method in the literature (Section 4).

2 Related Work

Word sense disambiguation (WSD) is an important research area in Natural Language Processing. Most methods (e.g., (Navigli, 2009; Ranjan Pal and Saha, 2015)) proposed over the years use the context of an ambiguous word occurrence to determine the intended sense.

A task more closely related to our work is disambiguating words in groups (e.g. a list of synonyms in a thesaurus) without context. These groups of words do not come with a context often used in typical WSD setting. In previous work, external lexical resources were often employed to make up for the lack of context. Dagan and Itai (1994) suggested resolving lexical ambiguities of the target language using statistical data from another language. Resnik (1999) proposed to disambiguate a related group of nouns using Wordnet hypernyms (Miller, 1995). Tuan et al. (2020) employed clustering the definition-based word sense vectors to filter out unrelated senses and obtain intended senses of the given group. Our method, which we will describe in detail in the next section, addresses a similar problem of assigning senses relevant in an academic context to a given much larger list of academic keywords.

In the research area of developing word lists for EAP and ESL teaching and learning, the best known work is West (1953)'s General Service List (GSL), a list consisting of approximately 2000 words to represent the most frequently-used part of the English vocabulary. Later, Coxhead (2000) compiled the Academic Word List (AWL), which places emphasis on words more frequently used in academic texts, excluding words in the GSL. It includes words that are observed in various academic disciplines with frequencies above a given threshold, but are relatively uncommon in other kinds of texts. Paquot (2010) comes up with a list that marks a departure from AWL. With words in GSL allowed, the Academic Keyword List (AKL) is compiled from various academic writing corpora of different disciplines. AKL words are selected based on comparative analysis of an academic corpus with a fiction corpus to find words that carry more academic importance and that are shared across academic disciplines.

In contrast to the WSD and EAP work described above, we exploit translation ngrams and clusters of sense-definition-based embeddings to identify intended, EAP-relevant sense. Although being relatively simple, preliminary evaluation shows the method yields satisfactory results. The results also offer the benefit of combining the resources of an EAP word list and existing dictionaries (e.g., Cam-

bridge learner dictionary) to generate more pedagogically useful EAP materials for as pointed out in (Paquot, 2010).

3 Finding intended senses of a word list

Using contexts from a running text to find the intended sense of words in EAP word list might not work very well, as existing WSD methods typically have low accuracy rates and typically require an annotated dataset for training. We exploit the nature of the EAP genre and word list: there are often more than one way to express the same concept. By computing sense to sense similarity based on translation and word embedding, and selecting the sense with the highest similarity based on translations, we propose a method for identifying intended senses for each word in an EAP word list. We now formally state the problem we are addressing:

Problem statement: Given a set of words W of a known application (in this case, academic use), and all senses $s_{w,i}$ of each word $w_i \in W$. Our goal is to identify the sense $s_{w,i'}$ for each word w_i such that $s_{w,i'}$ is the sense that is used to facilitate academic writing.

The steps to our solution is as follows:

propose-1 (verb)

DIFFICULTY: B2

to offer or suggest a possible plan or action for other people to consider

SYNONYMS: suggest-1, propound-1, submit-2, float-5

ANTONYMS: withdraw-1, deny-2, refuse-1

EXAMPLE SENTENCES:

- I **propose** that we wait until the budget has been announced before committing ourselves to any expenditure.
- He **proposed** dealing directly with the suppliers.

Figure 2: A example of the disambiguation result

In the final result, the correct sense of each word is shown along with its English definition, guide word, difficulty level, and example sentences (Figure 2).

As an example, consider the group of words: (propose, suggest, argue): Figure 3 shows the count of each translation. For *propose*, since the sense [“建議”, “提議”, “提出”] contains the translations ’ 建議’ and ’ 提議’, which both occurred twice, the sense is selected as the ’correct’ sense. Similarly, the sense [“提議”, “建議”] is selected for *suggest*. On the other hand, since all translations of *argue* only occur once within the group, the ’correct’ academic sense cannot be identified, and its first sense [“爭論”, “爭吵”, “爭辯”] is selected.

Algorithm 1: Algorithm for finding the most likely sense given a group of semantically similar words W, based on the number of occurrences of all translated senses

```
def FreqDisambiguate(W, S);
Input :
  W : A set of words to disambiguate;
  S : A set of translated senses for each word in W, where each sense of word w is s_i^w ∈ S, and each translated sense s_{i,t}^w ∈ s_i^w;
  F : A hash table with translated sense s_{w,i} as keys and their frequencies in W as values.
Output:
  D: A hash table with words as keys and 1 disambiguate sense for each word as values
for each w_i in W
    initialize maxSenseCnt
    initialize topSense
    for each s_i^w in s^w
        for each s_{i,t}^w in s_i^w
            if F[s_{i,t}^w] > maxSenseCnt then
                maxSenseCnt ← F[s_{i,t}^w]
                topSense ← s_{i,t}^w
    D[w_i] ← topSense
return D
```

4 Experimental results

We conducted two sets of experiments: the first experiment disambiguate word senses based on the frequencies of the translated senses, as in Algorithm 1. The second experiment employs definition embedding and clustering, which requires only lexical resources from one language, as opposed to our first

```
argue [['爭論', '爭吵', '爭辯'], ['論證', '說理', '辯論'], ['顯示出', '表明']]
propose [['建議', '提議', '提出'], ['提名', '推薦（某人）擔任某職（或參加某組織）'], ['求婚'], ['計劃',
'打算']]
suggest [['提議', '建議'], ['暗示', '間接表明', '意味著'], ['使想到', '使聯想到']]

{'爭論': 1, '爭吵': 1, '爭辯': 1, '論證': 1, '說理': 1, '辯論': 1, '顯示出': 1, '表明': 1, '建議': 2
, '提議': 2, '提出': 1, '提名': 1, '推薦（某人）擔任某職（或參加某組織）': 1, '求婚': 1, '計劃': 1, '打
算': 1, '暗示': 1, '間接表明': 1, '意味著': 1, '使想到': 1, '使聯想到': 1}

argue ['爭論', '爭吵', '爭辯']
propose ['建議', '提議', '提出']
suggest ['提議', '建議']
```

Figure 3: A sample of the disambiguation process and result: The first block displays all senses of each word. The second block displays counts of each translation. The third block shows the senses chosen.

experiment where translation in another language is needed.

4.1 The Vanilla Approach: Most Frequent Translation

In our experiment, we retrieve the Mandarin translation of each word from the online Cambridge Dictionary (Cambridge University Press, 2021). The Cambridge Dictionary is an ideal external resource for this task because it is compiled by experienced lexicographers and provide rich information useful for learners, including proficiency level, English definitions, guide words, example sentences, and translations in many L1 languages.

The AKL consists of 930 potential academic words and phrases, categorized into 5 groups: nouns, verbs, adjectives, adverbs, and others. Table 1 is a summary of word distribution of the original and the translated AKL. The number of unique words and their occurrences are greater than the numbers from the original AKL because each word or phrase is often provided with more than one translation. Since words on the AKL are semantically related, translations of the same senses tend to occur multiple times, which enable us to identify the "correct"academic sense for each word by choosing the sense whose translation has the highest occurrences among all senses.

4.2 The Monolingual Approach: Clustering with Definition Embedding

While our method based on Mandarin translation is effective, it requires translations of a second language to work. The method would not function when translations are not available, or when counting the occurrences of translated token is not as simple as those in Mandarin, for instance, languages that involves extensive inflections. In this case, a pretrained sense embedding model could be a valid alternative.

Word embeddings have exploded in use since Mikolov et al. (2013). The ability to represent words in a continuous vector space enables a whole new array of applications. In recent years, sentence encoding models (or "language models") that utilize embeddings as word representations have evolved to achieve state-of-the-art results, notably seq2seq (Sutskever et al., 2014), ELMo (Peters et al., 2018), the Transformer (Vaswani et al., 2017), and most recently BERT (Devlin et al., 2018). These models typically encode the input sequences with Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or self-attention into hidden states, and map the hidden states to an output sequences with a decoder. Although several studies (Kågebäck and Salomonsson, 2016; Raganato et al., 2017; Ahmed et al., 2018; Wiedemann et al., 2019; Huang et al., 2020) have used these models to perform WSD, these studies all focus on WSD within context, whereas our task concerns a special case where context is not available. In addition, individual senses are implicitly incorporated in the hidden states or embeddings, which poses challenges for interpretability and downstream tasks that specifically require embeddings for individual senses. Last but not least, these language models are expensive to train, which makes them unsuitable for our task at hand.

Bosc and Vincent (2018) proposed an alternate method for encoding embeddings, where embeddings are encoded from dictionary defi-

Part of speech	AKL entries	Unique trsl. words	Total trsl. words
NOUN	355	1785	2174
VERB	233	1158	1415
ADJECTIVE	180	798	954
ADVERB	87	300	347

Table 1: This table shows the number of AKL entries and number of words after translation. Total translated words include repeated words.

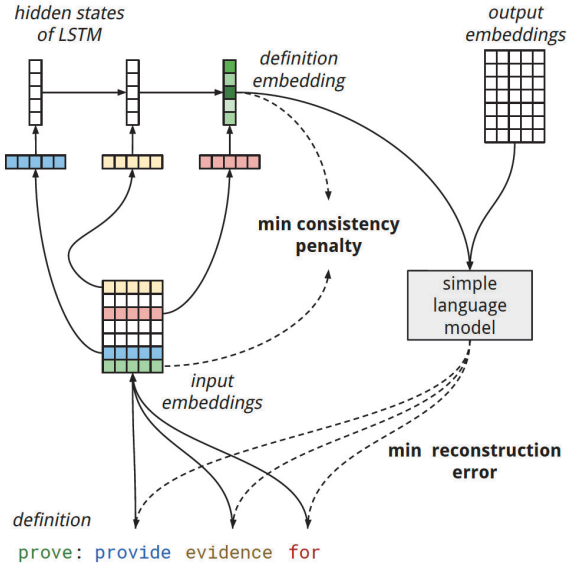


Figure 4: An overview of the CPAE model. Graph from Bosc and Vincent (2018)

nition. In Bosc and Vincent (2018), definition embeddings are encoded by passing the definition of a sense to an LSTM and a linear transformation. The authors also proposed a method called *consistency penalty*, where the objective function defined as the distance between the embedding produced by the LSTM and the embedding being fed into the LSTM is to be minimized. Fig. 4 provides an overview of the architecture of the model. In their final step, all definition embeddings of a word are concatenated into one embedding. For our purpose of extracting the correct sense from all other senses, we choose not to concatenate the definition embeddings and instead encode each definition as a separate embedding, enabling the actions for our next step. For our embeddings, definitions from Cambridge Dictionary are used as input data.

Based on the assumption that senses that are semantically related tend to be closer to each other in a vector space, we perform a clustering algorithm as in Tuan et al. (2020) on a complete graph derived from the sense embeddings of words in AKL:

1. Obtain pairwise similarity between all pairs of senses of all words. The pairwise similarity is calculated as the cosine similarity between two definition embeddings.
2. Build a complete graph with senses as nodes and their pairwise similarity as edges.
3. Perform density-based spatial clustering as in Campello et al. (2013) and obtain the largest cluster.
4. Senses within the largest cluster are selected as disambiguated senses.
5. Words whose senses are not included within the cluster are assigned the disambiguation result from the translation method.

5 Evaluation and Discussion

We randomly select 10 percent of identified senses from each category for evaluation. We have two human experts evaluate whether each sense identified is of academic use. For the baseline, the first sense listed for each word in Cambridge Dictionary is chosen as the "correct" sense. The accuracies of the translation method, the clustering method, and the baseline are summarized in Table 2.

Our method performs significantly better than the baseline, averaging at 90% accuracy, whereas the baseline has an average of 79% accuracy. The seemingly impossible 100% accuracy for adverbs is due to small group size, and the relatively lower degrees of sense ambiguity (1.9 senses per word vs. the average 2.6 senses per word).

part of speech	translation accuracy	clustering accuracy	baseline accuracy
NOUN	**91%**	**91%**	71%
VERB	**83%**	**83%**	75%
ADJECTIVE	**89%**	83%	83%
ADVERB	**100%**	**100%**	94%
AVERAGE	**90%**	87%	79%

Table 2: Accuracies of academic sense disambiguation

Interestingly, our evaluation also shows that using sense embeddings with clustering does not necessarily yield better results than the knowledge adn translation-based approach, although the embedding based approach also performs much better than the baseline. There could be several reasons for its lesser performance: One is that the sense embeddings might not have encompassed the complete semantic content for some senses, placing those senses at positions further away from the main cluster where other sense vectors are. On the other hand, the words grouped together in AKL might not necessarily be closely-enough related semantically to form a large cluster. Despite its lower accuracies, the clustering method still outperforms the baseline, and serves as an effective disambiguation method when sense translations of the target word groups are not available.

6 Conclusion and Future Work

We have introduced a method for disambiguating senses for academic usage for words in the Academic Keyword List. The method involves retrieving translations of each sense in another language and extracting English senses that correspond to the most frequently-occurring translation. We have also experimented with disambiguating senses with the clustering of sense embeddings. Both methods yields reasonable good results in disentagling academic senses from other senses. More importantly, our work marks a step towards building an academic writers' dictionary.

Many avenues exist for future research and improvement for our method. For building an academic writers' dictionary, the next step is to include all potentially-academic senses for each word on the AKL. Disambiguating AKL words within running text would largely benefit learners of EAP. Synonyms, antonyms, and example sentences of the disambiguated senses could be generated to further assist ESL learners.

Additionally, another direction of research would be to experiment on using translations in languages other than Mandarin for potentially better disambiguation results, as well as a more profound understanding of the properties of human sense-making.

Acknowledgments

The authors would like to thank the reviewers for their invaluable suggestions and feedback. Thanks also due to GLORIA Operation Center for offering a short course on Academic English Writing which provides the impetus of this research

References

Mahtab Ahmed, Muhammad Rifayat Samee, and Robert Mercer. 2018. A novel neural sequence model with multiple attentions for word sense disambiguation. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 687–694. IEEE.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium. Association for Computational Linguistics.

2021 Cambridge University Press. 2021. Cambridge dictionary. https://dictionary.cambridge.org. Accessed: 2021-08-01.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Averil Coxhead. 2000. A new academic word list. *TESOL Quarterly*, 34(2):213–238.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Comput. Linguist.*, 20(4):563—596.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2020. Glossbert: Bert for word sense disambiguation with gloss knowledge.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional lstm. *arXiv preprint arXiv:1606.03568*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39—41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Magali Paquot. 2010. *Academic vocabulary in learner writing: From extraction to analysis*. Bloomsbury Publishing.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.

Alok Ranjan Pal and Diganta Saha. 2015. Word sense disambiguation: A survey. *International Journal of Control Theory and Computer Modeling*, 5(3):1—16.

P. Resnik. 1999. *Disambiguating Noun Groupings with Respect to WordNet Senses*, pages 77–98. Springer Netherlands, Dordrecht.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Kai-Wen Tuan, Yi-Chien Lin, Jason S Chang, Kuan-Lin Lee, and Li-Kuang Chen. 2020. Consenses: Disambiguating content word groups based on knowledge base and definition embedding. In *2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 260–265. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Michael West. 1953. *A General Service List of English Words*. Addison-Wesley Longman Ltd, London.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings.