From Sense to Action: A Word-Action Disambiguation Task in NLP

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Yu-Hsiang Tseng
Graduate Institute of Linguistics
National Taiwan University
seantyh@gmail.com

Chiung-Yu Chiang
Graduate Institute of Linguistics
National Taiwan University
cychiang@ntu.edu.tw

Richard Lian
Graduate Institute of Networking and Multimedia
National Taiwan University
dclian@nlg.csie.ntu.edu.tw

Yong-fu Liao
Graduate Institute of Linguistics
National Taiwan University
mcku1115@gmail.com

Mao-Chang Ku
Graduate Institute of Linguistics
National Taiwan University
mcku1115@gmail.com

Ching-Fang Shih
Graduate Institute of Linguistics
National Taiwan University
r08142004@ntu.edu.tw

Abstract

Words are conventionalized symbols that present the function by which meaning is attached to form. The Word Sense Disambiguation, which has been taken as one of the core semantic processing tasks in the pipe-lined NLP architecture, aims to assign proper word sense to lemma form in varied contexts based on a word-sense inventory such as WordNet. However, there are some theoretical assumptions unattested from a functional linguistic point of view. This paper proposes an alternative by introducing a novel task called word action disambiguation task (WAD) concentrated on the observable pairs between words and actions. The accompanying dataset, which was manually edited and compiled, is composed of 419 multiple-choice questions. We further verified the dataset through item evaluation with human rating data, and the semantic relations among the dataset were annotated automatically. A baseline performance with an accuracy of 38.64% was also provided with BERT models and 43.18% after incorporating paradigmatic knowledge with semantic graph. We expect the proposal of the WAD task and dataset would motivate computational models to incorporate more complex aspects of human language.

1 Introduction

Due to its polysemous behavior, selecting the most appropriate sense for a word in a text has been one of the most important yet challenging NLP tasks over the years. Given a pre-defined sense inventory, computationally assigning each word in target texts with proper sense (thus Word Sense Disambiguation) is assumed to be crucial for MT, IR, QA, and other systems (Navigli, 2009). Although the sense inventory such as WordNet has been continuously maintained and implemented cross-linguistically, the issue regarding the extent to which the sense granularity (i.e., levels of semantic specificity) in the sense inventory would be sufficient for downstream NLP tasks remains less explored.

Three tacit and intertwined assumptions underlying the conventional WSD task are (1) word senses can be operationalized as discrete and distinguishable ones, (2) word senses (as included in the sense inventory) can be shared by the entire language community, and (3) WSD with the fine-grained sense specification can be successfully applied to actual language data, and facilitate a wide range of downstream NLP tasks. However, the reported poor inter-annotator agreement (IAA) and low reliability of sense distinction/annotation in the

task seem to falsify these assumptions and thus motivate projects like OntoNotes (Hovy et al., 2006; Cinková et al., 2012).

This paper aims to serve as a first attempt to propose an alternative to the underlying assumptions from the functional and granular linguistic perspective. First, the notion wordhood of as assumed in the WSD task is not self-evident, particularly for languages whose writing systems do not provide the delimiter of a word boundary. In this aspect, word segmentation or determination is rather theory-laden and would be best regarded as the wordhood annotation rather than the preprocessing task with ground truth as conventionally taken. Second, word-meaning pairs are fluid in nature, whose granularity (in terms of the length of the word and the functions it carries) is influenced by its underlying ontology (paradigmatic dimension), surrounding context (syntagmatic dimension) and real-world application (pragmatic force). Under this view, it is hard to get a common, static, or solid 'feel of sense' among native speakers. Finally, it is still unclear regarding the relation between WSD and Natural Language Understanding (NLU). For instance, what levels of granularity of sense (from fine-grained to coarse-grained) do we need for the machine comprehension, or in what sense can we justify that WSD is a sine qua non for NLU?

There has been a huge amount of related work trying to grapple with the WSD-related issues by exploiting various machine learning models (Navigli, 2009). On the resource side, in order to achieve better efficiency and performance, sense granularity in the sense inventory such as WordNet was explored and annotated in OntoNotes project (Weischedel et al., 2011; Palmer, Dang & Fellbaum, 2005). However, the paradigm underlying the WSD task has also been questioned since (Kilgariff, 1997), and sense discretization and enumerative view of word senses inventory that is implicitly/explicitly presume is strongly criticized as well (Pustejovsky, 1995). Consequently, we adopt a functional linguistic approach to the linguistic units and introduce the design of a novel task, which can be regarded as an in vivo evaluation of the WSD system.

In terms of language understanding, we see language as a communication device used to ask, demand, raise questions. The utterance, either in spoken or written forms, is an observable word sequence which encodes the speaker's illocutionary force, the "combination of the illocutionary point of an utterance, and the particular presuppositions and attitudes that must accompany that point" (Searle and Vanderveken, 1985). In pragmatics, illocutionary force further distinguishes the following types of acts: inquiring, promising, asserting, ordering, etc. As the words which serve as the building block in the sequence are mostly polysemous, it is thus commonly and naively assumed that one core part of our NLU competence depends on the identification of the correct word sense for each word in the utterance, and the understanding is accomplished in a compositional manner. That's the basic underlying philosophy of the current WSD task. However, these senses are unobservable theoretical constructs. In the communicative context, as long as listeners can react with proper observable responses, the mechanism underlying the word sense disambiguation inside the listener's mind is only latent constructs. That is, the listeners understand the utterances when they react with proper actions against them. This leads to Davidsonian notion of action (Davidson, 1985), that an action is something an agent does that was 'intentional under some description'. The relations among words, senses, and the Davidsonian actions with a framework is depicted in Figure 1.

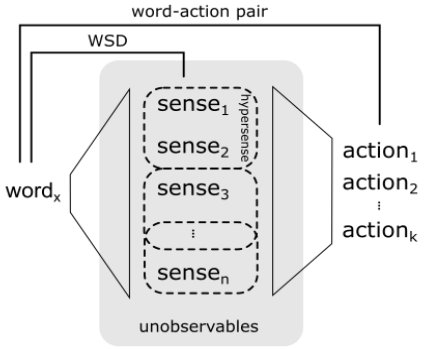

Figure 1: A schematic explanation of the relationships between words, word senses, and proposed *actions*.

To illustrate the relationship between words and actions, we develop a novel task called word action disambiguation task (WAD) to highlight the communication and dynamic aspect of word and action. The action inevitably reduced to textual descriptions in the task in order to be efficiently processed by machines. However, the proposed task underlines the interactions between words and actions by emphasizing the pragmatic and context-dependent

nature among them, and the relationship between them cannot be solely determined by lexical semantics.

2 Word Action Disambiguation Task and its Dataset

This section explains the proposed task and its corresponding dataset to alleviate the issues when splitting fine-grained, continuous word senses as assumed in previous WSD studies. The new task concentrates on the observable words and actions elicited. The task is implemented in the form like multiple-choice decision. A dataset with 439 items was also compiled to accompany the proposed task by 9 annotators. In each item, the question states a scenario, situation, or dialogue, in which a critical word is embedded. The critical words are polysemous single-character words selected from CWN. Resulting from the word's polysemy, four possible descriptions of actions are listed as options. An agent's (models or computer agents) task is to select the most proper action based on the understanding of the critical word's sense.

The critical words are selected from Chinese Wordnet (CWN). Followed by rigorous lexical-semantic theories, CWN distinguishes fine-grained differences between word senses. In the WAD dataset, we selected 400 single-character verbs with more than 3 verbal senses. Among these senses, we defined 4 critical senses of each word where proper action would be impossible if the word senses are conflated. For example, 叫 "jiao4" has 13 senses listed in CWN. In the sentence: “餓了就叫水餃來吃 (Order some dumplings if you are hungry.)”, the sense of the critical word 叫 (jiao4) refers to “order something”. If the agent misunderstands it as calling someone over, 水餃 (shui3jiao3) would be a human, not a kind of food, the resulting actions would be improper.

A complete WAD task item is as follows. We first identify 4 critical word senses and created multiple-choice questions and options (the critical word is marked with angle brackets):

我昨天<吃>了公館夜市的臭豆腐，真棒

I <had> stinky tofu in Gongguan Night Market yesterday. That was great!

A. 難怪假日的時候人潮都很多
   No wonder it is so crowded on weekends.
B. 做這事真耗體力，不划算
   It is not worthy of doing such labor-consuming work.
C. 這機器太爛了吧，卡插進去就拔不出來
   This machine sucks. You can’t get the card back after you insert it.
D. 貨物這麼重喔，難怪船無法停泊在這港口
   The cargo must be heavy. No wonder the cargo ship cannot anchor here.

The critical word, 吃 (chi1), has 28 senses in CWN. The question states a scenario in a night market, using the sense of 吃 (chi1) which refers to “eat something”. Options followed are 4 other possible responses toward based on other critical senses: (A) to eat something in; (B) to consume lots of resources; (C) to indicate that a card is captured by a cash machine, and (D) to displace the water while the boat is immersed in the sea. The correct answer to the question is option A.

These 4 options refer to the respective sense by the frame semantics, pragmatics, context, or common-sense knowledge. Importantly, the options are designed not to relate to the question with lexical semantics alone. That is, the questions and options are designed so the mapping relations between words and actions cannot be easily learned by models based on current syntagmatic vector semantics.

The proposed WAD dataset is aimed to be pragmatically, contextually, real-world relevant word action pairs, and these pairs cannot be determined by a model trained only on syntagmatic relations. Therefore, we verify the dataset with two approaches. (1) Item evaluation: we collect human raters' responses on these items and select the most appropriate items to include in the final dataset (Section 3). (2) Dataset Evaluation: we attempted a current deep learning model; the resulting performance is a tentative baseline on the proposed dataset (Section 4).

3 Item Evaluation

We evaluated items in the dataset with human ratings. Results of rating data were used to select the most appropriate items to include in the final dataset. We first describe methods of collecting rating data and item selection results (Section 3.2).

3.1 Item Rating Study

Five Mandarin native speakers, aged from 19 to 24, were recruited in the rating study. After researchers gave instructions, raters were asked to evaluate how well each option matches the question stem. We used a 5-point scale Likert scale on each rating item: from definitely not the correct answer (point 1), not likely to be the correct answer (point 2), possibly incorrect or possibly correct (point 3), likely to be the correct answer (point 4), and definitely the correct answer (point 5). Each rater went through all 1756 question-option pairs. They responded with independent spreadsheets so that ratings data would not be seen by other raters.

There were 8,780 rating scores collected. The mean and the standard deviation of each question-option pairing were shown in Figure 2. The rating means of each pair are bimodally distributed, where modes occurred in point 1 and point 5. The pattern was expected as it indicated the raters tend to agree on which option should or should not be the appropriate choice. The fact that the frequency of ratings with higher scores (above 4) is lower than the frequency of ratings with lower scores (below 2) also aligns with this expectation since only a quarter of the options were designed to be the appropriate choices in the sense-action dataset.

The standard deviation of the ratings for each option signified inter-rater agreements. If raters did not agree on a pair, the rating scores would differ widely, resulting in a large standard deviation; on the contrary, if raters all agree on a pair and gave it the same scores, the standard deviation would be 0. As shown in Figure 2, the distribution of the standard deviations is right-skewed, with most of the standard deviations (64%) having values below 1.0. This indicates a high agreement on the ratings among the raters.

3.2 Item Selection

We devised a two-phase selection scheme each employing a criterion to select appropriate items respectively: agreement criterion and contrast criterion. Two indices were calculated for each criterion: (1) Agreement between correct (as designated by the question authors) and maximally rated options and (2) the ratio between the highest and second-highest rating.

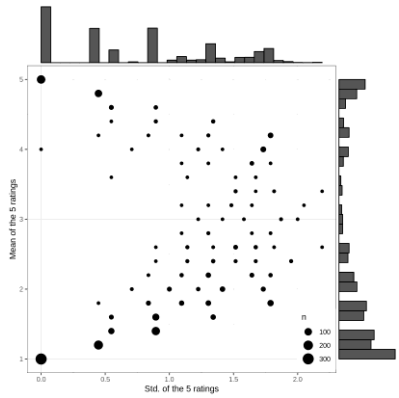

Figure 2: Distributions of the mean and standard deviation of the 5 raters' ratings for the options in the dataset

The agreement between correct and maximally rated options indicated the appropriateness of the answer created by the question authors. If the correct answer is rated lower than other options, the question was clearly not suitable in the dataset and therefore dropped. There were 10 questions omitted in this phase. This process filters out ten sense-action pairs and yields 429 remaining pairs (98%). In the second phase, we remove those pairs where the ratio between the highest and second highest below 1.15. This index indicated the ambiguity of the correct answer among other candid options. If the correct options were rated close to other options, the questions may involve complicated pragmatic or context considerations that cannot be resolved clearly even by human raters. There were further 10 items dropped in this phase. After two phases of item selection, there were 5% of dropped items and resulted in 419 items included in the final dataset[1].

4 Dataset Evaluation

WAD task involves learning the relations between words and actions, where pragmatic, semantic, and common-sense knowledge interact with each other. To evaluate the extent how current machine learning models perform on the WAD task, we compare two different models with two different feature representational approaches as baseline models of the dataset.

[1] Dataset is available at https://github.com/lopentu/WAD

4.1 Feature Representation

Two feature representation approaches are explored in this study. The first approach takes advantage of recent development of contextualized embedding models, specifically BERT (Devlin et al., 2018), to train a multiple-choice model on the proposed WAD task. Past studies showed that, as a transformer based model, a pre-trained BERT model is learned to represent lexical semantics of words and their syntactic relations within the sentences (Manning, Clark, Hewitt, Khandelwal, & Levy, 2020). This approach models the syntagmatic aspects of the linguistic inputs.

However, WAD items are designed to involve more than words' syntagmatic behaviors. Therefore, we devise a second approach to represent the information in items, which is more aimed to capture the paradigmatic relations among the stem and options in an item. Lexical resources, such as ConceptNet (Speer, Chin, & Havasi, 2017), is incorporated into the model through constructing a semantic graph. The graph has all the words in the dataset as nodes and relations (as defined in lexical resources) as edges. The resulting graph consists of 15,600 nodes and 807,426 edges. The graph contains 633 components (groups of nodes connected with each other), 608 of which are single node components. The largest component is composed of 12,469 nodes. An example of the semantic annotation on an item is shown in Figure 3. The semantic graph is further encoded into vectors with node embeddings (Grover & Leskovec, 2016). The hypothesis is that, equipped with paradigmatic and syntagmatic knowledge, the agent performs better in the WAD task.

4.2 Model Results

The dataset is split into a training set and a validation set with 80% and 20% proportions, respectively. A training example is composed of each of the four options concatenated with the question stem, resulting in a vector of four vectors, each one representing a question-option pair. The model needs to learn the indices of the correct answers.

Two models are trained and compared. The first model only uses BERT embedding as input, and a standard multiple-choice readout head, which is composed of a fully-connected layer of 768 hidden units, is stacked upon the output embeddings. The model finally predicts the index of the correct

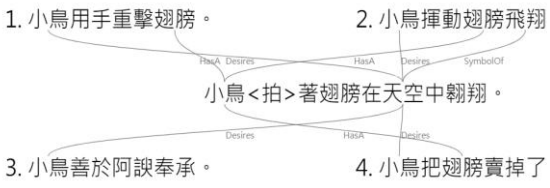

Figure 3: An example of an annotated item. The question stem is in the middle, surrounded by the four candidate options. The links among them are the semantic relations.

question-option pair. The second model's input includes BERT embeddings and node embeddings derived from the semantic graph. The input sequence of node embeddings is fed into a GRU layer, in which the hidden size is 100. The last hidden states of GRU are transformed with a fully-connected layer and concatenated with the BERT output as generated in the first model. We trained these two models on the WAD dataset with a batch size of 8, and the parameters are optimized with Adam optimizer with a learning rate of 5e-5 for 3 epochs.

The first model, with BERT embeddings only, achieved 38.64% accuracy, which was above randomly choosing (25% in a 4-option multiple-choice problem). The model with BERT and semantic graph embeddings achieves better performance with an accuracy of 43.18%. The pattern suggests paradigmatic information is helpful in learning the WAD task.

The current model with contextualized embeddings and semantic graph node embeddings can be considered as a tentative baseline performance for the WAD task. Distinctive from the traditional word sense disambiguation task, where word senses are mostly determined by its syntagmatic context, the WAD task deals further with pragmatics and real-world knowledge. These contextual knowledges are only implied in the text. The common-sense knowledge extracted from ConceptNet is a tentative approach that paves the way for a more comprehensive scheme. Such a scheme may involve annotating the common sense or real-world knowledge suggesting relations underlying the question stem and candidate options. Therefore, the connections between the question stem and correct options would be more accessible for a machine learner.

5 Conclusion

The enumerative and discretization of word senses impose profound limitations, both theoretically and computationally, on fine-grained sense inventories. In addition, the relationship between WSD and NLU remains unclear. Even given the success of WSD/sense tagger, how does that WSD process can logically entail the proper understanding of response in context? In this paper, we bring a 'meaning-in-action' philosophy into the WSD field. We identified the relations between words, senses, and actions and emphasize the observable pairs among them, i.e. word-action pairs. We then proposed a new task called "word-action disambiguation" (WAD), and its accompanying dataset which consisted of 419 multiple-choice questions. The task is designed to incorporate the semantic, pragmatic, real-world aspects of linguistic uses, and the relations between question and option pairs cannot be reduced to merely lexical semantics. We further evaluate each item with human rating data, to ensure the correctness and clearness of each item. A deep learning model, based on BERT, was trained on the WAD dataset to serve as a baseline performance. We expect the proposal of the WAD task and dataset would shed new light to the current architecture of WSD and motivate computational models to incorporate more complex aspects of human language.

Acknowledgement

This work was supported by Ministry of Science and Technology (MOST), Taiwan. Grant Number MOST. 108-2634-F-001-006.

References

Bojanowski, P., Grave, E, Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. Available on arXiv preprint arXiv:1607.04606.

Chang, L. L., Chen, K. J., & Huang, C. R. (2000). A Lexical-semantic Analysis of Mandarin Chinese Verbs: Representation and Methodology. *Computational Linguistics and Chinese Language Processing*, 5, 1-18.

Cinková, S., Martin Holub, Vincent Kríž (2012). Optimizing Semantic Granularity for NLP-report on a Lexicographic Experiment. In: *Proceedings of the 15th EURALEX International Congress*.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., & Hu, G. (2019). Pre-Training with Whole Word Masking for Chinese BERT. arXiv preprint arXiv:1906.08101.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Grover, A. & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).*

Hovy, E., M. Mitchell, M. Palmer, L. Ramshaw, and R. Weischedel (2006). 'OntoNotes: The 90% Solution'. In: *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers, NAACL-Short '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 57–60.

Kilgarriff, A. (1997). 'I Don't Believe in Word Senses.' *Computers and the Humanities* 31: 91– 113.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, Vol. 41, No. 2.

Palmer, Martha, Dang, Hoa & Fellbaum, Christiane. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* (13): 137-163.

Pustejovsky, J. (1995). The Generative Lexicon. MIT press.

Searle, J.R., Vanderveken D. (1985). Speech Acts and Illocutionary Logic. In: Vanderveken D. (eds) *Logic, Thought and Action. Logic, Epistemology, and the Unity of Science*, vol 2. Springer, Dordrecht.

Searle, J.R. (1990). Collective Intentions and Actions. In P. Cohen, J. Morgan, and M. Pollak (eds.). *Intentions in Communication*, Cambridge, MA: MIT Press.

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In Proceedings of AAAI 31.

Tripodi, R. & Pelillo, M. (2017). A Game-theoretic Approach to Word Sense Disambiguation. *Computational Linguistics*, 43(1):31-70.

Weischedel, R., et al. (2011). OntoNotes Release 4.0. Philadelphia: Linguistic Data Consortium.