Annotating Chinese Word Senses with English WordNet: A Practice on OntoNotes Chinese Sense Inventories

Hongzhi Xu[1], **Jingxia Lin**[2], **Sameer Pradhan**[3], **Mitchell Marcus**[3], **Ming Liu**[4]
[1]Shanghai Int. Studies Univ.; [2]Nanyang Tech. Univ.; [3]Univ. of Penn; [4]The HK Polytech. Univ.
hxu@shisu.edu.cn, jingxialin@ntu.edu.sg, pradhan@cemantix.org

Abstract

In this paper, we present our exploration of annotating Chinese word senses using English WordNet synsets, with examples extracted from OntoNotes Chinese sense inventories. Given a target word along with the example that contains it, the annotators select a WordNet synset that best describes the meaning of the target word in the context. The result demonstrates an inter-annotator agreement of 38% between two annotators. We delve into the instances of disagreement by comparing the two annotated synsets, including their positions within the WordNet hierarchy. The examination reveals intriguing patterns among closely related synsets, shedding light on similar concepts represented within the WordNet structure. The data offers as an indirect linking of Chinese word senses defined in OntoNotes Chinese sense inventories to WordNet sysnets, and thus promotes the value of the OntoNotes corpus. Compared to a direct linking of Chinese word senses to WordNet synsets, the example-based annotation has the merit of not being affected by inaccurate sense definitions and thus offers a new way of mapping WordNets of different languages. At the same time, the annotated data also serves as a valuable linguistic resource for exploring potential lexical differences between English and Chinese, with potential contributions to the broader understanding of cross-linguistic semantic mapping.

Keywords: OntoNotes, Lexical Resource, Chinese Word Senses, WordNet, Example-based Mapping

1. Introduction

NLP has entered the era of Large Language Models (LLMs). The most recent LLMs including GPT (Brown et al., 2020), LLaMA (Touvron et al., 2023) etc. have demonstrated remarkable cross-linguistic capabilities in various NLP tasks, including capturing language structures as well as implicitly learning and even generating world knowledge. Consequently, there is a growing trend within the NLP community to focus on cross-lingual applications. While LLMs have achieved outstanding performance in many challenging tasks, it is important to note that their training primarily relies on plain text data, without explicit modeling of syntactic and semantic information. Studies have shown that incorporating semantics can enhance the natural language understanding capabilities of these models (Levine et al., 2019; Blloshmi et al., 2020). However, constructing semantic resources, particularly those that are multilingual, is a complex and costly endeavor.

The Princeton WordNet (Miller et al., 1990) is one of the most widely used lexical semantic resources. Princeton WordNet (hence force, WordNet or simply PWN) has been used as a standard reference for English word sense disambiguation and proves invaluable for various NLP applications, including modeling of word similarities. Recognizing its significant value, extensive work has been devoted to creating WordNets in other languages and their mapping with PWN (Huang and Hsieh, 2010; Rudnicka et al., 2012; Strankale and Stade, 2022; Bakay et al., 2019; Cristea et al., 2004). These linked multilingual lexical resources will be potentially useful for various NLP tasks, e.g. cross-lingual word sense disambiguation (Lefever and Hoste, 2010), evaluating multi-lingual word embeddings, information retrieval, machine tranlatsion, question answering and other cross-lingual applications.

The OntoNotes project is a comprehensive initiative that spans various linguistic layers, including syntax, proposition, co-reference, and word sense and ontology (Hovy et al., 2006). It has played a pivotal role in the NLP field. The OntoNotes 5.0 (Weischedel et al., 2013) has been widely used for a diverse range of NLP tasks. The corpus consists of data in three languages: English, Chinese and Arabic. The English word senses are annotated with collapsed PWN synsets. This choice is driven by the fine-grained nature of WordNet word senses, which can make reliable agreement in word sense annotation challenging. However, for the Chinese data, different sense inventories are used, independent of their English counterparts. This linguistic disparity among the languages in OntoNotes presents a significant hurdle for the corpus's utilization in cross-lingual NLP applications. It's also worth noting that there are unassigned WordNet synset offsets for each defined Chinese word sense in the sense inventories, suggesting a potential future endeavor to establish links between Chinese word senses and their English counterparts.

In this paper, we attempt to fill this gap within

LREC-COLING 2024, pages 1187–1196

OntoNotes 5.0 by linking the word senses in Chinese sense inventories to PWN synsets. Due to the brevity and occasional vagueness of sense definitions in OntoNote 5.0, along with the presence of senses lacking any definitions, a direct mapping from senses to WordNet synsets can be challenging. To overcome this, we employ an example-based annotation approach.

We extract usage examples for each sense of a word defined in the Chinese sense inventories of OntoNotes 5.0 for annotation purposes. When dealing with a target word associated with a given example, annotators are tasked with identifying the most appropriate synset from the PWN for that target word. Our annotation results reveal an agreement of approximately 38% between two annotators. Beyond annotation discrepancies, the cases of disagreement shed light on intriguing semantic relationships among the selected WordNet synsets for the same target words.

To resolve the disagreements, a third annotator adjudicates by selecting one of the two options. Through a meticulous analysis, we uncover instances where examples from different senses (of either the same word, i.e. polysemy, or different words) in Chinese sense inventories are linked to the same English synsets (many-to-one mappings), while examples from the same sense in Chinese are occasionally annotated with different English synsets (one-to-many mappings). A comprehensive discussion of these findings is presented in Section 4. The data is freely available online and readers can conduct their own studies with the data tailored to specific research objectives and needs[1].

The remainder of this paper is organized as follows. In Section 2, we discuss previous work related to our study. Section 3 describes the details of the annotation including the guideline, the data, and the annotation result. Section 4 is our discussion about the annotation result. Section 5 offers our conclusions and outlines future avenues for our research.

2. Related Work

Our work generally follows the line of mapping Chinese word senses to PWN. PWN (Fellbaum, 1998) is a lexical database organized in a hierarchy (more strictly speaking, a graph) of concepts conveyed by sets of lexical items (synsets). The links in the hierarchy represent several semantic relations including hyponymy, meronymy, antonymy, etc. The lexicon mainly belongs to four parts of speech: nouns, verbs, adjectives, and adverbs. PWN has published different versions, with the most recent being Version 3.1, which will be used in this study.

Due to the significant value of PWN to NLP, considerable efforts have been devoted to creating similar WordNets in other languages and mapping them to PWN. The Euro WordNet project (Vossen, 1998) creates a multi-lingual database with several European languages including Dutch, Italian, Spanish, German, French, Czech and Estonian, all mapped to PWN. The inter-lingual linkage between various WordNets is facilitated through the so-called Inter-Lingual-Index, which is a unstructured list of concepts created from WordNet 1.5 synsets and associated with domain labels such as sports and military. Then all the synsets in different WordNets are linked with an equivalent relationship to one of the concepts in the list. During the mapping, if no suitable concept is found for a particular synset from the source WordNet, a new concept can be created and added to the Inter-Lingual-Index. When adding a new language, the only potential change is the augmentation of the Inter-Lingual-Index with the inclusion of new concepts.

This framework has motivated most of the subsequent work in mapping various languages to PWN, e.g. Chinese, Polish, Latvian, Turkish, and others (Rudnicka et al., 2012; Strankale and Stade, 2022; Bakay et al., 2019; Cristea et al., 2004; Huang and Hsieh, 2010). Bond and Foster (2013) make a notable attempt to link WordNets from over 80 languages using an automatic method that identifies corresponding WordNet synsets for the senses from the source WordNet by exploiting their English translations. The evaluation shows high precision in finding target WordNet synsets. While the method represents a promising initiative for creating linkings of WordNets, the recall rate remains an area of uncertainty.

In the case of Chinese, there are several available Chinese WordNet resources. One such resource is the Sinica Bilingual Ontology WordNet, known as Sinica BOW (Huang et al., 2004), which is developed through the translation of PWN synset words into Chinese using an English-Chinese dictionary. In other words, Sinica BOW assumes Chinese has a hierarchical structure similar to PWN. The Chinese WordNet (Huang and Hsieh, 2010)is constructed independently but follows a scheme akin to PWN. The synsets are manually created based on real examples and are subsequently linked to PWN synsets and the SUMO Ontology. Chinese Open WordNet (Wang and Bond, 2013) is created by integrating several Chinese WordNets with the consideration of low coverage of the Chinese WordNet (Bond and Foster, 2013).

In summary, most of the existing work of mapping WordNets of other languages to PWN are based

[1] https://github.com/xuhongzhi/ontonotes-sense-mapping

on synset mapping, which is mostly done through an automatic mapping method that heavily relies on word-to-word dictionary. The precision of such automatic mapping is not guaranteed. Also, translation of English WordNet cannot guarantee the recall rate of all possible words as well.

However, it's crucial to recognize that although English words can be translated into Chinese, these translations often fail to fully convey the nuanced meanings of their English counterparts (He, 2002; Pan, 2012; Guo and Song, 2020, among others). For instance, consider the English word 'serendipity', which embodies the unique connotation of unexpected, fortunate discoveries, a concept for which there is no direct equivalent in Chinese (the closest translation would be the noun phrase 意外收获 *yiwai shouhuo*) 'unexpected gain/profit'. Similarly, while 'parent' is commonly translated as 家长 *jiazhang* in Chinese, the latter term has a broader denotation encompassing not only parents but also other elder generations within a family, such as grandparents.

Moreover, significant differences in connotations abound. For instance, 'share', often translated as 分享 *fenxiang* in Chinese, conveys a neutral tone, whereas the latter is typically used exclusively for sharing joyful information or entities. These instances highlight how translating English words into Chinese may lack precision, and suggest the intricacies of aligning WordNets across different languages. The automatic mapping of synsets may not guarantee precision, and the translation of English WordNet may not ensure the inclusion of all possible words or senses, thus raising concerns about recall rates.

3. The Annotation Framework

There are several possible ways to link Chinese word senses to PWN synsets. The first is in line with most practices that directly map Chinese word senses to PWN synsets. However, it is important to note that the definitions of senses in Chinese sense inventories are typically concise, which can lead to potential issues with accurate interpretation by annotators and thus increase annotation errors.

Another way to perform the linking is to leverage existing Chinese WordNets that have been linked to WordNet. This approach simplifies the annotation process as annotators only need to choose from a limited set of senses for a given word in context. However, the effectiveness of this method largely depends on the quality of these pre-existing Chinese WordNets. As previously mentioned, most of these Chinese WordNets are automatically generated, and there is no guarantee that they compass all possible senses for a word. Furthermore, the linking errors existing in the current WordNets will be inherited and thus cause the issue of error accumulation.

Therefore, to address these potential limitations, we annotate individual Chinese words in context directly with PWN synsets. By letting annotators choose the best PWN synset for the given word, we can indirectly link the Chinese word senses to the English WordNet synsets. There are several merits of using the example-based method: 1) the annotation process is straightforward and can avoid inaccurate definitions of the original senses; 2) context information can be used for precise sense determination; 3) it can identify incorrect usage examples provided for particular senses; 4) it allows one-to-many mapping if multiple examples are provided for a single sense.

It's worth noting that the annotation process can be further facilitated by providing annotators with automatic suggestions of candidate WordNet synsets. These suggestions can be generated using current technologies based on Chinese-English dictionaries and pretrained language models. However, to avoid potential bias towards a specific model, it's advisable not to use automatic suggestions during the initial annotation. Nevertheless, the annotated data we provide can be instrumental in testing and refining automatic mapping methods.

3.1. The Data

In the Chinese sense inventories of OntoNotes 5.0, there are 762 Chinese words (lemmas) including verbs and nouns, saved in separate files. Each word file contains the corresponding senses of the word and each sense is defined with English meta language associated with usage examples in Chinese. In total, these lemmas collectively account for 2235 senses[2], among which 660 senses are empty (i.e. lack associated usage examples), 1113 senses are associated with a single example, and 462 senses include multiple examples. In Sum, there are 2240 examples in the inventories. The information is shown in Table 1. Additionally, the distribution of the number of senses for these lemmas is visually represented in Figure 1. Notably, the average number of senses for a lemma is approximately 2.9, with a standard deviation of 2.8. The lemma with the most senses contains a total of 40 distinct senses.

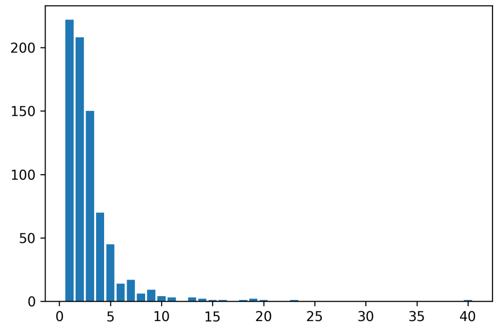
3.2. The annotation procedure

The annotation procedure can be described as follows: when presented with a target word along with an example sentence, the annotators try to identify and select the most appropriate synsets

[2]In OntoNotes senses, a dummy sense 'none of above' is placed for each word. We exclude all such dummy senses when counting the number.

#Lemmas	#Senses	#Examples
762	2235	2240

Table 1: The number of lemmas, number of senses, and number of examples in OntoNotes 5.0 Chinese sense inventories.



#Senses	1	2	3	4	5	≥ 6
#Lemmas	222	208	150	70	45	66

Figure 1: The distribution of the number of senses of lemmas in OntoNotes 5.0 Chinese sense inventories.

within WordNet. They are encouraged to employ a comprehensive set of English keywords or phrases to ensure they do not overlook any relevant synsets. In cases where multiple similar synsets appear to be suitable, annotators will look into the detailed definitions of these synsets, as well as their hypernym and hyponym to grasp the accurate comprehension of the underlying concept each synset represents, facilitating the selection of the most suitable one from the pool of candidates. In situations where no exact synsets align with the sense of the target word, annotators should opt for the most appropriate synset available, typically selecting a more general concept that encompasses the original ones.

During the annotation process, annotators are provided with only one piece of information: the target word highlighted within a given sentence. They are not given access to the original sense definitions from the OntoNotes sense inventories or any additional contextual information. This decision is made due to concerns about the reliability of the OntoNotes sense inventories. Consequently, annotators are not bound by specific parts of speech for the target words.

For our annotation, we choose WordNet 3.1, the most recent version of WordNet. We have also made it more convenient for annotators by providing access to an online query interface that facilitates searching for potential synsets using keywords or navigating the synset hierarchy [3].

Agreed	867 (38.7%)
Disagreed	1373 (61.3%)
Disagreed POS	244
Total	2240

Table 2: Annotation Agreement.

In our annotation effort, we have enlisted the assistance of ten annotators, all are master students from the English department of Shanghai International Studies University. Each example is evaluated by two annotators. In cases where there is a disagreement between these two annotators in selecting the appropriate synsets, a third annotator steps in to adjudicate and choose the most suitable synset from the two conflicting options.

4. Annotation Result and Discussion

In this section, we describe the annotation results in terms of annotation agreement, and will discuss instances where disagreements occurred. For cases where a sense is provided with multiple examples, we will examine if all the examples for the sense are consistently annotated with the same target synset. For those examples where the annotators disagree, we will also examine how the two chosen synsets are related to each other, e.g. how the two synsets are connected within the WordNet hierarchy, and whether the disagreements arise from errors or some interesting intrinsic similarities between the two synsets.

4.1. Example-level agreement

At the example level, we ignore the original senses of target words and focus on how many examples result in consensus between the two annotators regarding their chosen target synsets. The agreement result is shown in Table 2. Out of the total of 2,240 examples (target words), 867 (38.7%) are annotated with exactly the same target synsets by the two annotators, and the remaining 1373 examples are annotated with different synsets. Considering that annotators are tasked with choosing a single synset from a vast array of candidates within the entire WordNet, and they operate without any presupposed categories or information, this level of agreement is reasonable. Furthermore, we will demonstrate that the instances of disagreement often unveil intriguing semantic relationships between the synsets defined within WordNet.

Since Chinese lacks morphological cues to mark word categories, it is sometimes challenging to determine the category of a specific word, even within a given context. There are 244 instances where annotators choose different parts-

[3] https://wordnetweb.princeton.edu/

	adj	adv	noun	verb
adj	105	47	24	81
adv	0	40	5	40
noun	0	0	116	47
verb	0	0	0	868

Table 3: The confusion matrix of part-of-speech categories in annotation.

POS pair	Example
adj vs. adv	人才 / 流失 / 严重 talents loss serious ‘Talents are running away seriously.’
adj vs. noun	宽 / 与 / 严 lenience and strictness ‘lenience and strictness’
adj vs. verb	颜色 / 和 / 气质 / 接近 color and temperament close ‘The color and temperament are close.’
adv vs. verb	台湾 / 一直 / 与 / 日本 / 并列 / 为 / 帛琉 / 最 / 主要 / 的 / 旅客 / 来源 Taiwan always and Japan tie be Palau most main DE tourist source ‘Taiwan and Japan are both the main source of tourists of Palau.’
noun vs. verb	反对党 / 可能 / 会 / 因此 / 受到 / 指责 opposition-party maybe will thus receive accuse ‘The opposition party may thus be accused.’

Table 4: Examples of disagreed parts of speech.

of-speech for the target words, with the majority of discrepancies occurring between verbs and adjectives. Additionally, there are 1,129 examples where the annotators selected different synsets while still agreeing on the part of speech. Table 3 illustrates the instances of confusion between different parts of speech and some representative examples are shown in Table 4. Take the case of 接近 *jiejin* ‘close’ as an example, Chinese adjectives can directly follow subjects and act as predicates, which are typically verbs, causing the disagreement. There is indeed a discrimination of two different kinds of adjectives, predicative adjectives and non-predicative adjectives.

4.1.1. Disagreement Analyses

For the disagreed examples where annotators have selected different WordNet synsets, we take a closer look at the path connecting one of the chosen synset to the other within the WordNet hierarchy. This analysis reveals how the two synsets are semantically related and whether their differences are attributed to annotation errors or other factors, such as inherent similarities. Here, we exclude the cases where two synsets are not connected in the hierarchy. It is worth mentioning that adjectives and adverbs have limited hierarchical structures, and thus all 105 pairs of adjectival synsets and 40 pairs of adverbial synsets are excluded. Due to the over flattened structure of verbs in PWN, only 177 out of 868 pairs of verbal synsets are connected. Finally, there are 177 pairs of verbal synsets and 115 nominal synsets are kept. Particularly, we calculate the length of these paths between the synsets, and present the distribution of the path lengths in Figure 2. Notably, nouns tend to exhibit a higher average path length than verbs. This difference arises from the deeper hierarchy of nouns compared to verbs within WordNet.

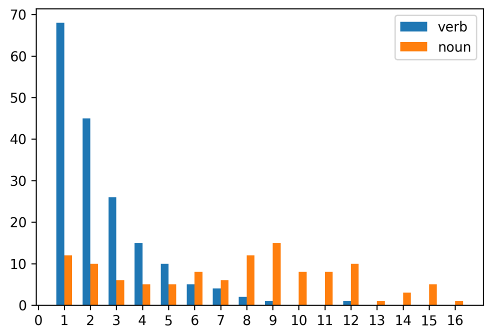


Figure 2: The distribution of distance between disagreed synsets. Synsets that do not connect each other (infinitive distance) are ignored, remaining 177 verbs and 115 nouns.

While certain instances of disagreement between annotators can be attributed to misinterpretations of the target words, a significant number of these disagreements actually reveal intriguing semantic relationships between the two selected synsets. These relationships often encompass diverse aspects of the same concepts, distinctions between literal and contextual meanings, and subtle nuances such as causative relationships.

In one particular category of disagreement, we observe that one synset serves as a hypernym of the other. An illustrative example is presented in Table 5. In this case, both the annotators make use of the keyword ‘continue’ to search for suitable synsets for 保留 *baoliu*, leading to the selection of two distinct synsets, one of which is a hypernym of the other.

Table 6 shows another example of disagreement,

巴伦西亚 / 是 / 西班牙 / 著名 / 的 / 农业区 / , / 至今 / 仍 / 保留 / 着 / 古老 / 的 / 阿拉伯式 / 灌溉 / 系统 / 。 Valencia be Spain famous DE agricultural-region , so-far still retain ZHE ancient DE Arabic irrigation system . 'Valencia is a famous agricultural region in Spain, it still retains an irrigation system of Ancient Arabic.'
original def.: to preserve (eg. culture, ruin, etc.)
Ann-1: {02415305:v:retain} allow to remain in a place or position or maintain a property or feature Ann-2: {02685709:v:continue} keep or maintain in unaltered condition; cause to remain or last Path: retain → prolong → continue

Table 5: An example of disagreement where one synset is a hypernym of the other, and the path from the synset to the other is two. The format of synset representation: {offset:pos:word}.

七 / 小时 / 后 / 又 / 被 / 新党 / 说服 / 而 / 打消 / 此 / 意 seven hour later again BEI Xin-Party persuade then give up this idea 'Seven hours later, he was persuaded by Xin party and give up the idea.'
original def.: thought, idea, opinion
Ann-1: {00164054:n:intention} an act of intending; a volition that you intend to carry out Path-1: intention → volition → choice → action → act → event → psychological feature Ann-2: {05991800:n:mind} your intention; what you intend to do Path-2: mind → purpose → goal → content → cognition → psychological feature

Table 6: An example of two annotated synsets Sharing the same hypernym. The paths from synset one and synset two to the shared hypernym are given with synset keywords with arrow.

where both synsets share the same hypernym. The length of the path between the two synsets is 9, indicating a substantial distinction between the two related concepts. From a semantic standpoint, these two synsets depict two facets of the same psychological state: the act of forming an intention and the resultant action intended to be executed.

We also find instances, where the choices made by both annotators are arguably correct, as demonstrated by the examples in Table 7. We can see from the pairs of the selected synsets that they are defined very similarly and no path is available between them.

浦东新区 / 制定 / 的 / 法规性 / 文件 / 有些 / 比较 / "/ 粗 / " Pudong formulate DE regulatory files some relative " coarse " 'Some regulatory files that Pudong new area formulated are relatively coarse.'
Original def.: coarse, crude, rough (as contrasted to fine, jing) Ann-1: {02237329:adj:crude} not carefully or expertly made Ann-2: {00919090:adj:approximate} not quite exact or correct
这 / 种 / 看法 / 是 / 不 / 对 / 的 this kind opinion be not correct DE 'This kind of opinion is not correct.'
Original Def.: correct, right Ann-1: {00636250:adj:correct} correct in opinion or judgment Ann-2: {00634232:adj:correct} free from error; especially conforming to fact or truth
吉尔吉斯 / 独立 / 以后 / 两 / 国 / 关系 / 发展 / 顺利 Kirghizia independent after two country relationship develop smooth 'After Kirghizia becomes independent, the relationship between the two countries develops smoothly.'
Original Def.: (adj) to be autonumous (politics, military) Ann-1: {00733338:adj:independent} free from external control and constraint Ann-2: {01066082:adj:autonomous} (of political bodies) not controlled by outside forces

Table 7: Examples where both of the two chosen synsets for the same target word are arguably correct.

4.2. Sense-level agreement

In the dataset, there are 462 senses associated with multiple examples. These senses offer an opportunity for us to examine if different examples from the same Chinese sense lead to a consensus on the same target synset, and if not, whether they suggest potential one-to-many mappings or inaccuracies within the provided examples. It should be noted that the sense-level agreement is inspected based on the final result after adjudication, namely each target word along with the usage example is associated with one single target synset.

Table 8 shows the results of our analysis for sense-level agreement. Among the 462 senses with multiple examples, only 119 senses have all of their examples annotated with the same target synset. The remaining 343 senses have examples anno-

single example senses	1113
multiple example senses	462
Agreed senses (one-to-one)	119
Disagreed senses (one-to-many)	343

Table 8: The sense-level agreement of the annotation.

#examples	#senses	agreed ratio
2	322	0.295
3	101	0.208
4	27	0.111
≥ 5	12	0.0
Total	462	0.258

Table 9: The agreement ratio of the original senses grouped by the number of usage examples. Here, the agreement means that all usage examples of the same sense are all annotated with the same target sense, yielding one-to-one mappings.

tated with different synsets.
Table 9 shows the ratio of senses in which all usage examples are consistently annotated with the same synset, which creates a one-to-one mapping between the original sense and the target synset. Among all the senses with two usage examples, only about 30% form one-to-one mappings. As the number of examples increases, the ratio of one-to-one mappings decreases accordingly. This indicates that as more examples are considered, the alignment between the original sense and the target synset becomes more nuanced and complex. This complexity may result from the various facets and contexts in which a word or concept can be used.

4.2.1. One-to-many mapping

The disagreed senses here refer to original senses that have multiple usage examples for which the target synsets are different, i.e. the so-called one-to-many mappings. An example is shown in Table 10. The two examples have the same original sense for the target word 吃 *chi* 'eat', but are annotated with different target synsets. A closer examination of the detailed definitions for these two synsets reveals a semantic relationship: they both pertain to the act of eating, albeit with a nuanced difference. One emphasizes the action of eating itself, while the other places the focus on the food being consumed. The subtle discrepancy between these two interpretations lies in whether the manner of eating or the food itself is emphasized. Additional context can potentially resolve this ambiguity. Nevertheless, it's essential to note that both interpretations are valid and contextually appropriate.

中国人 / 重 / 吃 Chinese emphasize eating 'Chinese people emphasize eating a lot.'
Synset-1 {00840028:n:eating}: the act of consuming food
身为 / 兄长 / 的 / 他 / , / 张罗 / 过 / 年 / 的 / 吃 / 穿 as elder-brother DE he , prepare celebrate Spring-Festival DE eating clothing 'As an elder brother, he is busy preparing the food and clothes for celebrating the Spring Festival. '
Synset-2 {00021445:n:food}: any substance that can be metabolized by an animal to give energy and build tissue

Table 10: An example of one-to-many mappings, namely two usage examples of the same original sense are annotated with two different target synsets.

Target synset {01781131:v:like}: find enjoyable or agreeable
爱 / 看 / 连续剧 / 的 / 阮金蓉 love watch TV-series DE Ruan-Jinrong 'Ruan Jinrong who loves watching TV series'
好事者 / 多 / 好 / 捶丸 meddler tend-to like Chui-wan (a soccer-like game) 'Meddlers tend to like Chuiwan as well.'

Table 11: An example of many-to-one mappings, namely senses of two different lemmas are mapped to the same target synset.

4.2.2. Many-to-one mapping

Many-to-one mappings refer to cases where different senses of either same (namely synonyms) or different words are mapped to the same target synset. In our dataset, we have identified 282 cases where different words share the same target synset. Many of these examples involve highly typical synonyms, such as 爱 *ai* 'love' and 好 *hao* 'like', 摆 *bai* 'lay' and 放 *fang* 'place', 包 *bao* 'wrap' and 裹 *guo* 'swaddle'. Interestingly, many of these word pairs can be combined to form a single compound word, e.g. 爱好 *ai-hao* 'like', 摆放 *bai-fang* 'lay out' 包裹 *bao-guo* 'swaddle'. The examples for 爱 *ai* 'love' and 好 *hao* 'like' are shown in Table 11. Sometimes, certain annotated synonyms overlook distinctions related to the situation aspect. Occasionally, words that represent activities and achievements/accomplishments are assigned to the same target synset, even though they exhibit different events and, thus, may behave differently in syntax. However, such nuanced semantic differences are often not explicitly encoded in WordNet. For instance, both 建 jian 'build' and 建成

jiancheng 'build-finish' share the core meaning of 'set up or build', but they differ in event structure. The former represents an activity, while the latter signifies an achievement event that entails a result.

Another type of many-to-one mappings involves different senses of the same word. For example, the word 保持 'keep' is given with three different senses in OntoNotes, i.e. 'continue', 'maintain' and 'remain'. However, it is notable that all examples associated with these senses are mapped to the same target synset. The details are shown in Table 12. A closer examination of the definitions for these three senses, in conjunction with their respective examples, reveals the challenge of distinguishing between them based solely on the Chinese examples. In English, 'continue', 'maintain' and 'remain' indeed convey nuanced differences. For instance, 'continue' describes the state's status and implies that it also existed before the speech time. In comparison, 'maintain' necessitates an agent subject and implies effort in sustaining the state. 'Remain' does not imply that the state existed before the speech time. However, the question of whether the Chinese word 保持 *baochi* 'keep' genuinely has such subtle distinctions requires special consideration. Notably, in the most widely used Chinese dictionary *xiandai hanyu cidian* "modern Chinese dictionary", the word is defined with a single sense, i.e. 保持 (原状), 使不消失或减弱 *baochi yuanzhuang, shi bu xiaoshi huo jianruo* 'keep the original state, make not to disappear or decrease'.

In this section, we exclusively explore several representative examples of each kind of disagreement, focusing on the semantic relationships between the two target synsets. We refrain from presenting comprehensive statistics for each situation because distinguishing between different types can be challenging, and this would necessitate separate annotations.

Ann: {02687605:v:keep} cause to continue in a certain state, position, or activity; e.g., 'keep clean'
Original Sense 1: to continue a certain state 贵州省 / 粮食 / 总 / 产量 / 已 / 连续 / 三 / 年 / 保持 / 在 / 一千万 / 吨 / 以上 Guizhou-province grain total output already continue three year keep at 10M ton above The grain output has been kept above 10M tons for three continuous years.
Original Sense 2: to maintain (a person's physical or psychological state,or social status) 我们 / 还是 / 保持 / 一 / 个 / 正面 / 肯定 / 的 / 一 / 个 / 态度 we still keep one CL positive positive DE one CL attitude 'We still keep a positive attitude.'
Original Sense 3: remain (silent) 中国 / 的 / 媒体 / 对 / 高行健 / 获奖 / 一 / 事 / 几乎 / 保持 / 沉默 China DE media towards Xingjian-Gao win-prize one event almost keep silence. 'Chinese media keep silence about the fact that Xingjian Gao won the prize.'

Table 12: An example of many-to-one mapping, for which the different senses are difficult to differentiate from the Chinese examples.

5. Conclusion and Future Work

In this paper, we present our practice of annotating Chinese word senses with PWN synsets. We use the usage examples extracted from OntoNotes Chinese sense inventories. The annotation results demonstrate a reasonable level of agreement, considering the challenge of the task. Our study offers an indirect mapping from the word senses defined in the Chinese sense inventories in OntoNotes 5.0 to PWN synsets, potentially enhancing the data's value for cross-lingual NLP tasks. Instead of providing a single definitive mapping, we have chosen to release the complete annotated dataset, which will enable researchers to create the final mapping according to their specific requirements. This can involve selecting the most prominent synsets for senses with multiple examples or retaining the one-to-many mapping results, among other possibilities. The data set will also serve as a valuable resource for linguistic studies, particularly in the context of exploring lexical relations between Chinese and English. Furthermore, this study provides insights into a novel sense mapping scheme that relies on individual examples rather than traditional sense definitions. It also provides guidance for developing models for automatic sense mapping when usage examples are available, contributing to expected performance improvements.

In future, we would like to explore the possibilities of automatic cross-lingual word sense alignment using state-of-the-art pre-trained multilingual language models and LLMs.

6. Limitations

In this study, we focused solely on senses within the OntoNotes Chinese sense inventories that feature at least one usage example. Consequently, the 660 senses lacking examples were not mapped to PWN synsets. In the future, our intent is to gather examples for these instances and conduct annotations to enhance the comprehensiveness of the dataset. Additionally, annotating supplementary examples for senses with a single usage instance would facilitate further investigations into one-to-many and many-to-one mappings.

7. Ethics Statement

All annotators in this study are full-time students at Shanghai International Studies University. They have been recruited as part-time research assistants and are compensated with standard wages in accordance with the university regulations.

8. Bibliographical References

Jordi Atserias, Luis Villarejo, and German Rigau. 2004. Spanish wordnet 1.6: Porting the spanish wordnet across princeton versions. In *LREC*.

Özge Bakay, Özlem Ergelen, and Olcay Taner Yıldız. 2019. Integrating turkish wordnet kenet to princeton wordnet: The case of one-to-many correspondences. In *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. Xl-amr: Enabling cross-lingual amr parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marine Carpuat. 2013. Nrc: A machine translation approach to cross-lingual word sense disambiguation (semeval-2013 task 10). In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 188–192.

Hsin-Hsi Chen, Chi-Ching Lin, and Wen-Cheng Lin. 2002. Building a chinese-english wordnet for translingual applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(2):103–122.

Dan Cristea, Catalin Mihaila, Corina Forascu, Diana Trandabat, Maria Husarciuc, Gabriela Haja, and Oana Postolache. 2004. Mapping princeton wordnet synsets onto romanian wordnet synsets. *Romanian Journal on Information Science and Technology, Special Issue on Balka-Net*, 7:125–145.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Hongjie Guo and Dan Song. 2020. *Jiyu Yuliaoku de Yinghan Duibi yu Fanyi (Corpus-based English-Chinese contrast and translation studies)*. Shanghai: Fudan University Press.

Shanfen He. 2002. Contrastive studies of english and chinese languages.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Chu-Ren Huang. 2003. Sinica bow: integrating bilingual wordnet and sumo ontology. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 825–826. IEEE.

Chu-Ren Huang, Ru-Yng Chang, and Hshiang-Pin Lee. 2004. Sinica bow (bilingual ontological wordnet): Integration of bilingual wordnet and sumo. In *LREC*.

Chu-Ren Huang and Shu-Kai Hsieh. 2010. Infrastructure for cross-lingual knowledge representation-towards multilingualism in linguistic studies. *Taiwan NSC-granted Research Project (NSC 96-2411-H-003-061-MY3)*.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2023. Overview of the trec 2022 neuclir track. *arXiv preprint arXiv:2304.12367*.

Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *5th International workshop on Semantic Evaluation (SemEval 2010)*, pages 15–20. Association for Computational Linguistics (ACL).

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. 2023. Synthetic cross-language information retrieval training data. *arXiv preprint arXiv:2305.00331*.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Roberto Navigli, Edoardo Barba, Simone Conia, and Rexhina Blloshmi. 2022. A tour of explicit multilingual semantics: Word sense disambiguation, semantic role labeling and semantic parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 35–43.

Petya Osenova and Kiril Simov. 2018. The data-driven bulgarian wordnet: Btbwn. *Cognitive Studies| Études cognitives*, (18).

Wenguo Pan. 2012. *Hanying Yuyan Duibi Gailun (An introduction to Chinese-English contrast)*. Beijing: Commercial Press.

Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. 2021. Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation. In *IJCAI*, pages 3915–3921.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A strategy of mapping polish wordnet onto princeton wordnet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048.

Laine Strankale and Madara Stade. 2022. Automatic word sense mapping from princeton wordnet to latvian wordnet. In *ICAART (1)*, pages 478–485.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *arXiv preprint arXiv:2210.09984*.

Xingran Zhu. 2020. Cross-lingual word sense disambiguation using mbert embeddings with syntactic dependencies. *arXiv preprint arXiv:2012.05300*.

9. Language Resource References

Ralph Weischedel and Martha Palmer and Mitchell Marcus and Eduard Hovy and Sameer Pradhan and Lance Ramshaw and Nianwen Xue and Ann Taylor and Jeff Kaufman and Michelle Franchini and Mohammed El-Bachouti and Robert Belvin and Ann Houston. 2013. *OntoNotes Release 5.0, LDC2013T19*. Philadelphia: Linguistic Data Consortium, ISLRN 151-738-649-048-2.