A Corpus of Adpositional Supersenses for Mandarin Chinese

Abstract

Adpositions are frequent markers of semantic relations, but they are highly ambiguous and vary significantly from language to language. Moreover, there is a dearth of annotated corpora for investigating the cross-linguistic variation of adposition semantics, or for building multilingual disambiguation systems. This paper presents a corpus in which all adpositions have been semantically annotated in Mandarin Chinese; to the best of our knowledge, this is the first Chinese corpus to be broadly annotated with adposition semantics. Our approach adapts a framework that defined a general set of supersenses according to ostensibly language-independent semantic criteria, though its development focused primarily on English prepositions (Schneider et al., 2018). We find that the supersense categories are well-suited to Chinese adpositions despite syntactic differences from English. On a Mandarin translation of The Little Prince, we achieve high inter-annotator agreement and analyze semantic correspondences of adposition tokens in bitext.

Introduction

Adpositions (i.e. prepositions and postpositions) include some of the most frequent words in languages like Chinese and English, and help convey a myriad of semantic relations of space, time, causality, possession, and other domains of meaning. They are also a persistent thorn in the side of second language learners owing to their extreme idiosyncrasy BIBREF1, BIBREF2. For instance, the English word in has no exact parallel in another language; rather, for purposes of translation, its many different usages cluster differently depending on the second language. Semantically annotated corpora of adpositions in multiple languages, including parallel data, would facilitate broader empirical study of adposition variation than is possible today, and could also contribute to NLP applications such as

machine translation BIBREF3, BIBREF4, BIBREF5, BIBREF6, BIBREF7, BIBREF8, BIBREF9 and grammatical error correction BIBREF1, BIBREF10, BIBREF11, BIBREF12, BIBREF13, BIBREF14.

This paper describes the first corpus with broad-coverage annotation of adpositions in Chinese. For this corpus we have adapted schneider-etal-2018-comprehensive Semantic Network of Adposition and Case Supersenses annotation scheme (SNACS; see sec:snacs) to Chinese. Though other languages were taken into consideration in designing SNACS, no serious annotation effort has been undertaken to confirm empirically that it generalizes to other languages. After developing new guidelines for syntactic phenomena in Chinese (subsec:adpositioncriteria), we apply the SNACS supersenses to a translation of The Little Prince (3 2 3), finding the supersenses to be robust and achieving high inter-annotator agreement (sec:corpus-annotation). We analyze the distribution of adpositions and supersenses in the corpus, and compare to adposition behavior in a separate English corpus (see sec:corpus-analysis). We also examine the predictions of a part-of-speech tagger in relation to our criteria for annotation targets (sec:adpositionidentification). The annotated corpus and the Chinese guidelines for SNACS will be made freely available online.

Related Work

To date, most wide-coverage semantic annotation of prepositions has been dictionary-based, taking a word sense disambiguation perspective BIBREF16, BIBREF17, BIBREF18. BIBREF19 proposed a supersense-based (unlexicalized) semantic annotation scheme which would be applied to all tokens of prepositions in English text. We adopt a revised version of the approach, known as SNACS (see sec:snacs). Previous SNACS annotation efforts have been mostly focused on English—particularly STREUSLE BIBREF20, BIBREF0, the semantically annotated corpus of reviews from the English Web Treebank BIBREF21. We present the first adaptation of SNACS for Chinese by annotating an entire Chinese translation of The Little Prince.

Related Work ::: Chinese Adpositions and Roles

In the computational literature for Chinese, apart from some focused studies (e.g., BIBREF22 on logical-semantic representation of temporal adpositions), there has been little work addressing adpositions specifically. Most previous semantic projects for Mandarin Chinese focused on content words and did not directly annotate the semantic relations signaled by functions words such as prepositions BIBREF23, BIBREF24, BIBREF25, BIBREF26. For example, in Chinese PropBank, BIBREF27 argued that the head word and its part of speech are clearly informative for labeling the semantic role of a phrase, but the preposition is not always the most informative element. BIBREF28 annotated the Tsinghua Corpus BIBREF29 from People's Daily where the content words were selected as the headwords, i.e., the object is the headword of the prepositional phrase. In these prepositional phrases, the nominal headwords were labeled with one of the 59 semantic relations (e.g. Location, LocationIni, Kernel word) whereas the prepositions and postpositions were respectively labeled with syntactic relations Preposition and LocationPreposition. Similarly, in Semantic Dependency Relations (SDR, BIBREF30, BIBREF31), prepositions and localizers were labeled as semantic markers mPrep and mRange, whereas semantic roles, e.g., Location, Patient, are assigned to the governed nominal phrases.

BIBREF32 compared PropBank parsing performance on Chinese and English, and showed that four Chinese prepositions (4, 2, 3, and 4) are among the top 20 lexicalized syntactic head words in Chinese PropBank, bridging the connections between verbs and their arguments. The high frequency of prepositions as head words in PropBank reflects their importance in context. However, very few annotation scheme attempted to directly label the semantics of these adposition words.

BIBREF33 is the most relevant adposition annotation effort, categorizing Chinese prepositions into 66 types of senses grouped by lexical items. However, these lexicalized semantic categories are constrained to a given language and a closed set of adpositions. For semantic labeling of Chinese adpositions in a

multilingual context, we turn to the SNACS framework, described below.

Related Work ::: SNACS: Adposition Supersenses

BIBREF0 proposed the Semantic Network of Adposition and Case Supersenses (SNACS), a hierarchical inventory of 50 semantic labels, i.e., supersenses, that characterize the use of adpositions, as shown in fig:supersenses. Since the meaning of adpositions is highly affected by the context, SNACS can help distinguish different usages of adpositions. For instance, single-label presents an example of the supersense Topic for the adposition about which emphasizes the subject matter of urbanization that the speaker discussed. In single-label-amb, however, the same preposition about takes a measurement in the context, expressing an approximation.

. I gave a presentation about:Topic urbanization.

. We have about:Approximator 3 eggs left.

Though assigning a single label to each adposition can help capture its lexical contribution to the sentence meaning as well as disambiguate its uses in different scenarios, the canonical lexical semantics of adpositions are often stretched to fit the needs of the scene in actual language use.

. I care about:StimulusTopic you.

For instance, eg:stimulustopic blends the domains of emotion (principally reflected in care, which licenses a Stimulus), and cognition (principally reflected in about, which often marks non-emotional Topics). Thus, SNACS incorporates the construal analysis BIBREF34 wherein the lexical semantic contribution of an adposition (its function) is distinguished and may diverge from the underlying relation in the surrounding

context (its scene role). Construal is notated by SceneRoleFunction, as StimulusTopic in eg:stimulustopic.

Another motivation for incorporating the construal analysis, as pointed out by BIBREF34, is its capability to adapt the English-centric supersense labels to other languages, which is the main contribution of this paper. The construal analysis can give us insights into the similarities and differences of function and scene roles of adpositions across languages.

Adposition Criteria in Mandarin Chinese

Our first challenge is to determine which tokens qualify as adpositions in Mandarin Chinese and merit supersense annotations. The English SNACS guidelines (we use version 2.3) broadly define the set of SNACS annotation targets to include canonical prepositions (taking an noun phrase (NP) complement) and their subordinating (clausal complement) uses. Possessives, intransitive particles, and certain uses of the infinitive marker to are also included BIBREF35.

In Chinese, the difficulty lies in two areas, which we discuss below. Firstly, prepositional words are widely attested. However, since no overt derivational morphology occurs on these prepositional tokens (previously referred to as coverbs), we need to filter non-prepositional uses of these words. Secondly, post-nominal particles, i.e., localizers, though not always considered adpositions in Chinese, deliver rich semantic information.

Adposition Criteria in Mandarin Chinese ::: Coverbs

Tokens that are considered generic prepositions can co-occur with the main predicate of the clause and introduce an NP argument to the clause BIBREF36 as in zho:shangtopic. These tokens are referred to as

coverbs. In some cases, coverbs can also occur as the main predicate. For example, the coverb 4 heads the predicate phrase in zho:pred.

. 1 4:Locus 24 4:TopicLocus 3342.

3sg p:at academia lc:on-top-of successful

`He succeeded in academia.'

. 3 4 de 2 4 4 34.

2sg want de sheep res at inside

`The sheep you wanted is in the box.' (zh_lpp_1943.92)

In this project, we only annotate coverbs when they do not function as the main predicate in the sentence, echoing the view that coverbs modify events introduced by the predicates, rather than establishing multiple events in a clause BIBREF37. Therefore, lexical items such as 4 are annotated when functioning as a modifier as in zho:shangtopic, but not when as the main predicate as in zho:pred.

Adposition Criteria in Mandarin Chinese ::: Localizers

Localizers are words that follow a noun phrase to refine its semantic relation. For example, 4 in zho:shangtopic denotes a contextual meaning, `in a particular area,' whereas the co-occurring coverb 4 only conveys a generic location. It is unclear whether localizers are syntactically postpositions, but we annotate all localizers because of their semantic significance. Though coverbs frequently co-occur with

localizers and the combination of coverbs and localizers is very productive, there is no strong evidence to suggest that they are circumpositions. As a result, we treat them as separate targets for SNACS annotation: for example, 4 and 4 receive Locus and TopicLocus respectively in zho:shangtopic.

Setting aside the syntactic controversies of coverbs and localizers in Mandarin Chinese, we regard both of them as adpositions that merit supersense annotations. As in zho:shangtopic, both the coverb 4 and the localizer 4 surround an NP argument 24 (`academia') and they as a whole modify the main predicate 3342 (`successful'). In this paper, we take the stance that coverbs co-occur with the main predicate and precede an NP, whereas localizers follow a noun phrase and add semantic information to the clause.

Corpus Annotation

We chose to annotate the novella The Little Prince because it has been translated into hundreds of languages and dialects, which enables comparisons of linguistic phenomena across languages on bitexts. This is the first Chinese corpus to undergo SNACS annotation. Ongoing adpositional supersense projects on The Little Prince include English, German, French, and Korean. In addition, The Little Prince has received large attention from other semantic frameworks and corpora, including the English BIBREF38 and Chinese BIBREF26 AMR corpora.

Corpus Annotation ::: Preprocessing

We use the same Chinese translation of The Little Prince as the Chinese AMR corpus BIBREF26, which is also sentence-aligned with the English AMR corpus BIBREF38. These bitext annotations in multiple languages and annotation semantic frameworks can facilitate cross-framework comparisons.

Prior to supersense annotation, we conducted the following preprocessing steps in order to identify the

adposition targets that merit supersense annotation.

Corpus Annotation ::: Preprocessing ::: Tokenization

After automatic tokenization using Jieba, we conducted manual corrections to ensure that all potential adpositions occur as separate tokens, closely following the Chinese Penn Treebank segmentation guidelines BIBREF39. The final corpus includes all 27 chapters of The Little Prince, with a total of 20k tokens.

Corpus Annotation ::: Preprocessing ::: Adposition Targets

All annotators jointly identified adposition targets according to the criteria discussed in subsec:adpositioncriteria. Manual identification of adpositions was necessary as an automatic POS tagger was found unsuitable for our criteria (sec:adpositionidentification).

Corpus Annotation ::: Preprocessing ::: Data Format

Though parsing is not essential to this annotation project, we ran the StanfordNLP BIBREF40 dependency parser to obtain POS tags and dependency trees. These are stored alongside supersense annotations in the CoNLL-U-Lex format BIBREF41, BIBREF0. CoNLL-U-Lex extends the CoNLL-U format used by the Universal Dependencies BIBREF42 project to add additional columns for lexical semantic annotations.

Corpus Annotation ::: Reliability of Annotation

The corpus is jointly annotated by three native Mandarin Chinese speakers, all of whom have received

advanced training in theoretical and computational linguistics. Supersense labeling was performed cooperatively by 3 annotators for 25% (235/933) of the adposition targets, and for the remainder, independently by the 3 annotators, followed by cooperative adjudication. Annotation was conducted in two phases, and therefore we present two inter-annotator agreement studies to demonstrate the reproducibility of SNACS and the reliability of the adapted scheme for Chinese.

tab:iaa-results shows raw agreement and Cohen's kappa across three annotators computed by averaging three pairwise comparisons. Agreement levels on scene role, function, and full construal are high for both phases, attesting to the validity of the annotation framework in Chinese. However, there is a slight decrease from Phase 1 to Phase 2, possibly due to the seven newly attested adpositions in Phase 2 and the 1-year interval between the two annotation phases.

Corpus Analysis

Our corpus contains 933 manually identified adpositions. Of these, 70 distinct adpositions, 28 distinct scene roles, 26 distinct functions, and 41 distinct full construals are attested in annotation. Full statistics of token and type frequencies are shown in tab:stats. This section presents the most frequent adpositions in Mandarin Chinese, as well as quantitative and qualitative comparisons of scene roles, functions, and construals between Chinese and English annotations.

Corpus Analysis ::: Adpositions in Chinese

We analyze semantic and distributional properties of adpositions in Mandarin Chinese. The top 5 most frequent prepositions and postpositions are shown in tab:statstoptoks. Prepositions include canonical

adpositions such as 14 and coverbs such as 4. Postpositions are localizers such as 4 and 1. We observe that prepositions 4 and 4 are dominant in the corpus (greater than 10%). Other top adpositions are distributed quite evenly between prepositions and postpositions. On the low end, 27 out of the 70 attested adposition types occur only once in the corpus.

Corpus Analysis ::: Supersense & Construal Distributions in Chinese versus English

The distribution of scene role and function types in Chinese and English reflects the differences and similarities of adposition semantics in both languages. In tab:statssupersensezhen we compare this corpus with the largest English adposition supersense corpus, STREUSLE version 4.1 BIBREF0, which consists of web reviews. We note that the Chinese corpus is proportionally smaller than the English one in terms of token and adposition counts. Moreover, there are fewer scene role, function and construal types attested in Chinese. The proportion of construals in which the scene role differs from the function (scene$\ne $fxn) is also halved in Chinese. In this section, we delve into comparisons regarding scene roles, functions, and full construals between the two corpora both quantitatively and qualitatively.

Corpus Analysis ::: Supersense & Construal Distributions in Chinese versus English ::: Overall Distribution of Supersenses

fig:barscenezhen,fig:barfunctionzhen present the top 10 scene roles and functions in Mandarin Chinese and their distributions in English. It is worth noting that since more scene role and function types are attested in the larger STREUSLE dataset, the percentages of these supersenses in English are in general lower than the ones in Chinese.

There are a few observations in these distributions that are of particular interest. For some of the examples, we use an annotated subset of the English Little Prince corpus for qualitative comparisons, whereas all quantitative results in English refer to the larger STREUSLE corpus of English Web Treebank reviews BIBREF0.

Corpus Analysis ::: Supersense & Construal Distributions in Chinese versus English ::: Fewer Adpositions in Chinese

As shown in tab:statssupersensezhen, the percentage of adposition targets over tokens in Chinese is only half of that in English. This is due to the fact that Chinese has a stronger preference to convey semantic information via verbal or nominal forms. Examples eg:enmoreadpositions,eg:zhlessadpositions show that the prepositions used in English, of and in, are translated as copula verbs (4) and progressives (44) in Chinese. Corresponding to fig:barscenezhen,fig:barfunctionzhen, the proportion of the supersense label Topic in English is higher than that in Chinese; and similarly, the supersense label Identity is not attested in Chinese for either scene role or function.

. It was a picture of:Topic a boa constrictor in:Manner the act of:Identity swallowing an animal . (en_lpp_1943.3)

. [4 de] 4 [[4 2 32] 44 12 [4 1 4 34]]

draw de cop one cl boa prog swallow one cl big animal

`The drawing is a boa swallowing a big animal'. (en_lpp_1943.3)

Corpus Analysis ::: Supersense & Construal Distributions in Chinese versus English ::: Larger Proportion

of Locus in Chinese

In both fig:barscenezhen and fig:barfunctionzhen, the percentages of Locus as scene role and function are twice that of the English corpus respectively. This corresponds to the fact that fewer supersense types occur in Mandarin Chinese than in English. As a result, generic locative and temporal adpositions, as well as adpositions tied to thematic roles, have larger proportions in Chinese than in English.

Corpus Analysis ::: Supersense & Construal Distributions in Chinese versus English ::: Experiencer as Function in Chinese

Despite the fact that there are fewer supersense types attested in Chinese, Experiencer as a function is specific to Chinese as it does not have any prototypical adpositions in English BIBREF35. In eg:enexperiencergoal, the scene role Experiencer is expressed through the preposition to and construed as Goal, which highlights the abstract destination of the `air of truth'. This reflects the basic meaning of to, which denotes a path towards a goal BIBREF43. In contrast, the lexicalized combination of the preposition 4 and the localizer 21 in eg:zhexperiencershenghuo are a characteristic way to introduce the mental state of the experiencer, denoting the meaning `to someone's regard'. The high frequency of 21 and the semantic role of Experiencer (6.3%) underscore its status as a prototypical adposition usage in Chinese.

. To:ExperiencerGoal those who understand life, that would have given a much greater air of truth to my story. (en_lpp_1943.185)

. [4:Experiencer [32 12 de 2] 21:Experiencer], 44 1 4 32 12

p:to know-about life de people lc:one's-regard this-way tell res seems real

`It looks real to those who know about life.' (zh_lpp_1943.185)

Corpus Analysis ::: Supersense & Construal Distributions in Chinese versus English ::: Divergence of Functions across Languages

Among all possible types of construals between scene role and function, here we are only concerned with construals where the scene role differs from the function (scene$\ne $fxn). The basis of hwang-etal-2017-double construal analysis is that a scene role is construed as a function to express the contexual meaning of the adposition that is different from its lexical one. fig:barconstrualzhen presents the top 10 divergent (scene$\ne $fxn) construals in Chinese and their corresponding proportions in English. Strikingly fewer types of construals are formed in Chinese. Nevertheless, Chinese is replete with RecipientDirection adpositions, which constitute nearly half of the construals.

The 2 adpositions annotated with RecipientDirection are 4 and 4, both meaning `towards' in Chinese. In eg:enrecipient,eg:zhrecipientdirection, both English to and Chinese 4 have Recipient as the scene role. In eg:enrecipient, Goal is labelled as the function of to because it indicates the completion of the “saying” event. In Chinese, 4 has the function label Direction provided that 4 highlights the orientation of the message uttered by the speaker as in eg:zhrecipientdirection. Even though they express the same scene role in the parallel corpus, their lexical semantics still requires them to have different functions in English versus Chinese.

. You would have to say to:RecipientGoal them: “I saw a house that costs $$20,000$.” (en_lpp_1943.172).

. (3) 41 [4:RecipientDirection 1men] 1: “3 44 le 2 4 24 32 de 2zi.”

2sg must P:to 3pl say 1sg see asp one CL $10,000$ franc de house

`You must tell them: “I see a house that costs 10,000 francs.” ' (zh_lpp_1943.172).

Corpus Analysis ::: Supersense & Construal Distributions in Chinese versus English ::: New Construals in Chinese

Similar to the distinction between RecipientGoal and RecipientDirection in English versus Chinese, language-specific lexical semantics contribute to unique construals in Chinese, i.e. semantic uses of adpositions that are unattested in the STREUSLE corpus. Six construals are newly attested in the Chinese corpus:

[noitemsep,topsep=0pt]

BeneficiaryExperiencer

CircumstanceTime

PartPortionLocus

TopicLocus

CircumstanceAccompanier

DurationInstrument

Of these new construals, BeneficiaryExperiencer has the highest frequency in the corpus. The novelty of this construal lies in the possibility of Experiencer as function in Chinese, shown by the parallel examples in eg:enbenibeni,eg:zhbeniexpe, where 4 receives the construal annotation BeneficiaryExperiencer.

. One must not hold it against:Beneficiary them . (en_lpp_1943.180)

. 33zimen 4:BeneficiaryExperiencer 42men 41 14 xie

children P:to adults should lenient comp

`Children should not hold it against adults.' (zh_lpp_1943.180)

Similarly, other new construals in Chinese resulted from the lexical meaning of the adpositions that are not equivalent to those in English. For instance, the combination of 1 ... 2 (during the time of) denotes the circumstance of an event that is grounded by the time (2) of the event. Different lexical semantics of adpositions necessarily creates new construals when adapting the same supersense scheme into a new language, inducing newly found associations between scene and function roles of these adpositions. Fortunately, though combinations of scene and function require innovation when adapting SNACS into Chinese, the 50 supersense labels are sufficient to account for the semantic diversity of adpositions in the corpus.

POS Tagging of Adposition Targets

We conduct a post-annotation comparison between manually identified adposition targets and

automatically POS-tagged adpositions in the Chinese SNACS corpus. Among the 933 manually identified adposition targets that merit supersense annotation, only 385 (41.3%) are tagged as adp (adposition) by StanfordNLP BIBREF40. fig:piegoldpos shows that gold targets are more frequently tagged as verb than adp in automatic parses, as well as a small portion that are tagged as noun. The inclusion of targets with pos=verb reflects our discussion in subsec:adpositioncriteria that coverbs co-occurring with a main predicate are included in our annotation. The automatic POS tagger also wrongly predicts some non-coverb adpositions, such as 12, to be verbs.

The StanfordNLP POS tagger also suffers from low precision (72.6%). Most false positives resulted from the discrepancies in adposition criteria between theoretical studies on Chinese adpositions and the tagset used in Universal Dependencies (UD) corpora such as the Chinese-GSD corpus. For instance, the Chinese-GSD corpus considers subordinating conjunctions (such as 23, 24, 42, 34) adpositions; however, theoretical research on Chinese adpositions such as BIBREF44 differentiates them from adpositions, since they can never syntactically precede a noun phrase.

Hence, further SNACS annotation and disambiguation efforts on Chinese adpositions cannot rely on the StanfordNLP adp category to identify annotation targets. Since adpositions mostly belong to a closed set of tokens, we apply a simple rule to identify all attested adpositions which are not functioning as the main predicate of a sentence, i.e., not the root of the dependency tree. As shown in Table TABREF43, our heuristic results in an F_1 of 82.4%, outperforming the strategy of using the StanfordNLP POS tagger.

Conclusion

In this paper, we presented the first corpus annotated with adposition supersenses in Mandarin Chinese. The corpus is a valuable resource for examining similarities and differences between adpositions in different languages with parallel corpora and can further support automatic disambiguation of adpositions

in Chinese. We intend to annotate additional genres—including native (non-translated) Chinese and learner corpora—in order to more fully capture the semantic behavior of adpositions in Chinese as compared to other languages.

Acknowledgements

We thank anonymous reviewers for their feedback. This research was supported in part by NSF award IIS-1812778 and grant 2016375 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.