Universal Semantic Tagging for English and Mandarin Chinese

Wenxi Li[1], **Yiyang Hou**[1], **Yajie Ye**[2], **Li Liang**[1*] and **Weiwei Sun**[3]
[1]Department of Chinese Language and Literature, Peking University
[2]Wangxuan Institute of Computer Technology, Peking University
[3]Department of Computer Science and Technology, University of Cambridge
{liwenxi,yiyang.hou,liangli15,yeyajie}@pku.edu.cn
ws390@cam.ac.uk

Abstract

Universal Semantic Tagging aims to provide lightweight unified analysis for all languages at the word level. Though the proposed annotation scheme is conceptually promising, the feasibility is only examined in four Indo–European languages. This paper is concerned with extending the annotation scheme to handle Mandarin Chinese and empirically study the plausibility of unifying meaning representations for multiple languages. We discuss a set of language-specific semantic phenomena, propose new annotation specifications and build a richly annotated corpus. The corpus consists of 1100 English–Chinese parallel sentences, where compositional semantic analysis is available for English, and another 1000 Chinese sentences which has enriched syntactic analysis. By means of the new annotations, we also evaluate a series of neural tagging models to gauge how successful semantic tagging can be: accuracies of 92.7% and 94.6% are obtained for Chinese and English respectively. The English tagging performance is remarkably better than the state-of-the-art by 7.7%.

1 Introduction

Developing meaning representations across different languages plays a fundamental and essential role in multilingual natural language processing, and is attracting more and more research interests (Costa-jussà et al., 2020). Existing approaches can be roughly divided into three categories: the crosslingual[1] approach focuses on lending semantic annotation of a resource-rich language, such as English, to an under-resourced language (Wang et al., 2019; Blloshmi et al., 2020; Mohiuddin and Joty, 2020); the interlingual approach attempts to provide a unified semantic framework for all languages (Abend and Rappoport, 2013; White et al., 2016; Ranta et al., 2020); the multilingual approach aims at developing comparable but not necessarily identical annotation schemes shared by different languages (Bond and Foster, 2013; Baker and Ellsworth, 2017; Pires et al., 2019).

In line with the interlingual approach, Universal Semantic Tagging (UST; Bjerva et al., 2016) develops a set of language-neutral tags (hereafter referred to as *sem-tag*) to annotate individual words, providing shallow yet effective semantic information. Semantic analyses of different languages utilise a same core tag set, but may also employ a few language-specific tags. Figure 1 presents an example.

English	*I*/PRO *had*/PST *repaired*/EXT *my*/HAS *watch*/CON ./NIL
German	*Ich*/PRO *hatte*/PST *meine*/HAS *Armbanduhr*/CON *repariert*/EXT ./NIL
Italian	*Ho*/NOW *riparito*/EXT *il*/DEF *mio*/HAS *orologio*/CON ./NIL
Chinese	我/PRO 把/OBJ 我/PRO 的/MOD 手 表/CON 修/EXT 好/EXT 了/PFT 。/NIL

Figure 1: An example of parallel sentences and their sem-tags. PRO: anaphoric & deictic pronouns; PST: past tense; EXT: untensed perfect; HAS: possessive pronoun; CON: concept; NOW: present tense; DEF: definite; OBJ: object; MOD: modification; PFT: perfect tense; NIL: empty semantics. All tags are universal, with the exception of non-core tags OBJ and MOD, which are newly created to annotate Chinese-specific linguistic phenomena that can not be represented by the existing system.

The idea of sem-tag is first applied to the Parallel Meaning Bank (PMB; Abzianidze et al., 2017), where a multilingual corpus, including Dutch, German and Italian, is semi-automatically built by projecting semantic tags from English sentences to

*This author is now working in Tencent.

[1] The terminology in the literature is quite diverse—the usages of "crosslingual", "interlingual" and "multilingual" vary from author to author.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5554–5566
June 6–11, 2021.

their translated counterparts. However, it is insufficient to prove the feasibility of UST only through some cases of inflectional and genetically related languages, because one main challenge in developing interlingual meaning representations is unifying annotations related to different characteristics of different languages. We argue that two questions with regard to universality of UST are still unanswered. Firstly, homologous words in PMB languages facilitate the application of UST, but it is not clear whether UST is equally applicable to languages sharing little cognates, although UST employs a delexicalised method. Another concern is from typology: it still remains unknown whether word-level semantic tags are effective for annotating long "sentence-words" composing many morphemes which are common in agglutinative languages (e.g. Turkish and Japanese) and polysynthetic languages (e.g. Eskimo languages).

This paper takes Mandarin Chinese, a phylogenetically distant language from the Indo–European family, as an example to explore the effectiveness of UST as a universal annotation scheme. Considering the balance of Chinese-specific linguistic properties and universality, we present a more comprehensive tag set where six new tags are added, indicating most sem-tags are applicable to Chinese (§2). Based on the new tag set, we establish a parallel corpus by manually translating WSJ into corresponding Chinese sentences and annotating sem-tags for 1100 sentence pairs. It is a peer-reviewed corpus with 92.9% and 91.2% inter-annotator observed agreement of Chinese and English respectively (§3). This relatively successful practice of UST in Chinese suggests it keeps the balance between the depth of represented information and the breadth of its coverage of languages. In other words, shallow semantics of UST enables it to be extended to annotate diversified languages.

By means of the newly created corpus, we evaluate a series of neural sequence labeling techniques (§4). The results demonstrate that the proposed scheme is promising with the accuracy of Chinese achieving 92.7% and the accuracy of English 94.6% (§5). The English tagging performance is remarkably better than the state-of-the-art (Abzianidze and Bos, 2017) by 7.7%, even though the sentences in our corpus are much longer than PMB on average, with 25 tokens per sentence compared with 6 in PMB.

In order to analyse the divergence between annotations of English and Chinese data and the plausibility of developing universal semantic representation in general, we manually annotate word alignment for 500 sentences. By studying the aligned counterparts, we argue that universality is still threatened to some extent because there are 37.0% aligned tokens with mismatched sem-tags. This phenomenon is mainly due to grammatical divergence, information loss of translation and difference of annotation strategies. All the analyses based on word alignment suggest that even for a delexicalised, relatively shallow meaning representation scheme, it can still be problematic to ensure that semantic representations could be comparable in a word-to-word way.

2 Tailoring Tag Sets for Mandarin Chinese

Considering different linguistic ways to encode tense, aspect, prepositions, measure words, subordinate clauses and comparative expressions, we provide a tailored version of UST to handle Mandarin Chinese. We present the complete tailored tag set in the Appendix.

Events and tense/aspect Different from English as well as many other Indo–European languages, there are no inflection-style tense-markers in Mandarin. Therefore, the morphological tense-related labels, e.g. ENS and EPS, are removed. Alternatively, temporal interpretation of Chinese can be conveyed through function words, adverbials or shared understanding of the context in Chinese (Smith and Erbaugh, 2005). Apart from the last way, the previous two are encoded by sem-tags FUT and IST. As for aspect in Chinese, there are only four commonly recognized aspect markers, denoting the preceding verbs are actualized or ongoing—了/过 are perfective (PFT) and 在/着 are progressive (PRG) (Liu, 2015).

Preposition Prepositions of English and Chinese vary in their history origins though they have similar syntactic function at present. English prepositions are mainly created to replace the lost inflectional case markers (Mitchell, 1985). On the other hand, Chinese prepositions can be traced to verbs. Li and Thompson (1989) even go so far as to call them *coverbs* since some of them are like verbs and can be used as verbs that have similar meanings. This term can avoid labeling them either verbs or prepositions. In this regard, Chi-

English	
EXS	untensed simple: *to walk, is eaten*
ENS	present simple: *we walk, he walks*
EPS	past simple: *ate, went*
EXG	untensed progressive: *is running*
EXT	untensed perfect: *has eaten*
Chinese	
EXS	untensed simple: 走、跑、休息
EXG	untensed progressive: 吃着、在看
EXT	untensed perfect: 换了、见过

Table 1: EVE tags of English and Chinese.

English	
NOW	present tense: *is sking, do ski, has skied, now*
PST	past tense: *was baked, had gone, did go*
FUT	future tense: *will, shall*
PRG	progressive: *has been being treated*
PFT	perfect: *has been going/done*
Chinese	
NOW	present tense: 现在
FUT	future tense: 将
PRG	progressive: 在，着
PFT	perfect: 了、过

Table 2: TNS tags of English and Chinese.

nese prepositions should not follow the practice on English because REL emphasizes grammatical relations between verbs and nouns while in Chinese the degree of grammarization of prepositions is not so far.

Consequently, we design a separate set of sem-tags for Chinesee prepositions by borrowing existing sem-tags (DXT/DXP/ALT) and adding some new sem-tags (MAN/RES/AIM/OBJ/COM).

Meaning	Sem-tag	Example
time & places	DXT / DXP	从、到、在、朝
manners	MAN	按照、用、被
reason & aim	RES / AIM	因、由于、为了
object	OBJ	对、和、替、连
comparative	COM	比、较
alternative	ALT	除、除了、除去

Table 3: Classification of Chinese prepositions and their corresponding sem-tags and examples.

Classifier Classifier is a Chinese-specific word class which is inserted between numerals and nouns to denote quantity. This category does not exist in English so we generalize UOM over the unit of measurement since its function is quite similar to classifiers (Li and Thompson, 1989).

Subordinate clause Whether subordinate clauses exist in Chinese is controversial since not all the clauses meet the standard *in a lower position than the main clause*. Additionally, words corresponding to subordinate conjunctions of English such as 因为 (because), 虽然 (although), etc, constitute a heterogeneous group and do not necessarily select a *subordinating* clausal complement (Paul, 2016). Given these two reasons, SUB is (temporarily) removed to avoid controversy.

Comparative expression UST designs a detailed label set to annotate comparative expressions in English. See Table 4. In particular, though expressions labeled as MOR/TOP and LES/BOT utilize exactly the same syntactic constructions, they are separated according to their meaning, in a way that is more oriented by applications. Different from English, Mandarin does not have morphological comparatives and superlatives. To express comparative-related meaning, adverbs 更 (roughly means *more*) and 最 (roughly means *most*) are utilized and annotated as MOR and TOP respectively. Accordingly, LES and BOT are deleted.

English	
EQU	equative: *as tall as John, whales are mammals*
MOR	comparative positive: *smarter, more*
LES	comparative negative: *less, worse*
TOP	superlative positive: *smartest, most*
BOT	superlative negative: *worst, least*
ORD	ordinal: *1st, 3rd, third*
Chinese	
EQU	equative: 这么、这样、和他一样高
MOR	comparative positive: 更
TOP	superlative positive: 最
ORD	ordinal: 第一、首次

Table 4: COM tags of English and Chinese.

3 The Corpus

We introduce a new moderate-sized corpus containing high-quality manual annotations for English and Chinese, which is now available at https://github.com/pkucoli/UST.

3.1 Data Source

To support fine-grained cross-lingual comparisons, the corpus includes 1100 parallel sentence pairs. We select 1100 sentences from the Wall Street Journal (WSJ) section of Penn TreeBank (PTB; Marcus et al., 1993). We choose it because it contains detailed semantic annotations and the sentences are relatively long, thus potentially carrying more complex information. It is noteworthy that various syntactic and semantic analyses of these English sentences have been built by multiple projects, e.g. DeepBank (Flickinger et al., 2012), PropBank (Palmer et al., 2005) and OntoNotes (Weischedel et al., 2013).

We then obtain Chinese counterparts of original English sentences by employing English–Chinese bilinguals to do literal translation. In addition, we also select 1000 sentences from Chinese TreeBank (CTB; Xue et al., 2005), where manual syntactic analyses are available.

3.2 Annotation

One doctoral student and one undergraduate student, majoring in linguistics, annotate the pair sentences. The guideline for English annotation is derived from the universal semantic tag set (Abzianidze and Bos, 2017) with reference to data in PMB and Chinese is annotated based on the modified tag set in the appendix. The annotation process consists of three steps: firstly, annotators independently annotate 100 Chinese WSJ sentences, and later compare and discuss disagreements between the annotations. The conflicting cases are then analyzed to modify the specification. After some iterations, the consistency between annotators is significantly improved. Additionally, we find part-of-speech (POS) tags are quite useful to accelerate manual annotation. Therefore, we apply the Stanford CoreNLP tool (Manning et al., 2014a) to get automatically predicted POS tags for the translated Chinese sentences.

Quality of the corpus The observed inter-annotator agreement in annotating Chinese and English sub-corpus data achieves 92.9% and 91.2% for Chinese and English sentences respectively. A high consistency in the annotation of both sub-corpus is obtained, which, in our view, demonstrates that UST is feasible for Chinese and the adjustment of original tag set is relatively satisfactory.

Re-tagging In order to improve the quality of annotation, we leverage the re-tagging strategy (Ide and Pustejovsky, 2017). Specifically, we investigate disagreements between initial model predictions and manual tagging, and correct manual annotation errors. After a round of re-tagging and re-training, the disagreement between the gold and the output of the tagger reduces from 10.3% to 7.9% on Chinese and 6.7% to 5.2% for English.

3.3 Divergence between English and Chinese Annotations

As a multilingual annotation scheme, UST represents semantic information in an interlingual way. Therefore, we want to answer after the modification of tag set, how the retained cross-lingual syntax and semantic divergence between distant languages still threatens its universality. We leverage a token-level word alignment for 500 parallel sentence pairs and investigate sem-tag mismatching between aligned tokens. Of the total 7295 pairs of tokens aligned, tokens in 3392 pairs share matched semantic tags with their counterparts, with a matching rate of **46.5%**. Note that punctuation and tokens tagged with NIL are excluded. Figure 2 shows an example of word alignment and sem-tag matching.

Our divergence analysis based on alignment is under the assumption that, as both the tasks of alignment and sem-tagging are concerning token-level semantic representation, the matched token pairs are expected to share the same sem-tags. Non-correspondence between aligned counterparts would therefore suggest divergence between the annotations in two languages, and further, may reveal problems caused by cross-lingual divergence.

Word alignment Word alignment between sentence pairs is firstly automatically acquired with Berkeley Aligner[2] and then manually corrected.

Matching rate and mismatches In general, aligned tokens are mostly entities or events, and among matches, the most frequent sem-tag is CON, followed by ORG and ROL. Other tags whose proportions in all matches exceed 3% are EXS, QUC, IST, PER and GPE. And the match per edge rates of these tags are also relatively high except for IST (see Table 5). However, since the mismatch phenomenon in CON, ORG and EXS are also

[2] https://code.google.com/archive/p/berkeleyaligner/

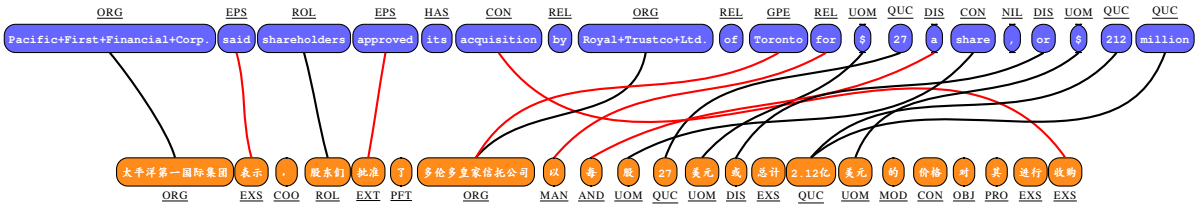


Figure 2: An example of alignment. Red lines shows that some aligned words may have different tags. ORG: organization; EPS: past tense; EXS: untensed simple; ROL: role; COO: coordination; EXT: untensed perfect; PFT: perfect; HAS: possessive pronoun; CON: concept; REL: relation; MAN: manner; AND: conjunction & univ. quantif.; UOM: unit of measurement; GPE: geo-political entity; DIS: disjunction & exist. quantif.; QUC concrete quantity NIL: empty semantics; MOD: modification; OBJ: object; PRO: modification.

not rare, annotation divergence could probably exist. A linguistically-motivated analysis suggests the following important factors:

- Grammatical divergence: an example is EXS in Figure 2. As illustrated in §2, it is used to tag Chinese verbs that are non-progressive or non-perfect, while only limited to untensed simple for English. This grammatical difference leads to tag set modification and thus results in sem-tag mismatch.
- Information loss caused by non-literal translation: In the example in Figure 2, *approved its acquisition* is translated as 批准...对其进行收购, which cause mismatch between *acquisition* (noun, CON) and 收购 (verb, EXS).
- Different annotation strategy for MWE: *Corp.* is tagged ORG while in their Chinese counterparts 公司 are tagged CON.

Sem-tag	Frequency	Correspondence
CON	34.9%	76.0%
ORG	8.7%	69.0%
ROL	7.2%	78.1%
EXS	6.0%	73,3%
QUC	5.8%	65.7%
IST	5.5%	31.7%
PER	4.4%	92.8%
GPE	4.4%	81.1%

Table 5: Frequency and correspondence rate of 8 sem-tags.

4 Tagging Models

Long Short-Term Memory (Hochreiter and Schmidhuber, 1997, LSTM) models have been widely used in various sequential tagging tasks (Huang et al., 2015; Ma and Hovy, 2016; Bohnet et al., 2018) and have achieved the state-of-the-art performance for many popular benchmark datasets. In our paper, we use Bidirecational LSTM (BiLSTM) with and without a Conditional Random Field (CRF) inference layer to build baseline systems for our dataset. In the rest part of this section, we will briefly formulate our baseline tagging models and introduce some widely used techniques that may enhance prediction for some tagging tasks.

Model For a word w_i in an input sentence $(w_1, w_2, ..., w_n)$, we use dynamically learned word embeddings $\boldsymbol{e}_1$ summed with the feature vectors calculated by BERT/ELMo after a linear projection $\boldsymbol{W}_e$ as the input of BiLSTM. If the POS tag of word w_i is used as additional input, we extend $\boldsymbol{x}_i$ with the the embedding $\boldsymbol{p}_i$ of the POS tag before passing it into the BiLSTM.

$$\boldsymbol{x}_i = \boldsymbol{e}_i + \text{BERT}(w_1, ..., w_n)_i \boldsymbol{W}_e$$

$$\boldsymbol{f}_i, \boldsymbol{b}_i = \text{BiLSTM}(\boldsymbol{x}_i \oplus \boldsymbol{p}_i, ...\boldsymbol{x}_n \oplus \boldsymbol{p_n})_i$$

After obtaining the contextual representations $\boldsymbol{f}_i$ and $\boldsymbol{b}_i$, we pass the concatenation of $\boldsymbol{f}_i$ and $\boldsymbol{b}_i$ to a multilayer perceptron (MLP) to calculate the scores vector $\boldsymbol{s}_i$ over semantic tags.

$$\boldsymbol{s}_i = \text{MLP}(\boldsymbol{f_i} \oplus \boldsymbol{b_i})$$

Finally, we feed $\boldsymbol{s}_i$ into a softmax layer to choose a tag with highest probability for each word independently, or a CRF layer which can select the tag sequence with highest probability for the whole sentence.

Subword/Character-level Models In order to solve the out-of-vocabulary (OOV) issues in sequence tagging tasks, many subword-level and character-level models are proposed (Akbik et al., 2018; Ling et al., 2015; Bohnet et al., 2018). We do not use these models for experiments, instead we leverage pretrained language models to handle OOV issues, such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018). These pretrained language models are trained on large corpus and use a subword/character-level vocabulary, which provide better contextual word representations.

POS features POS categories can provide low-level syntax infomation which is beneficial for sem-tagging. In our experiments, we try to use POS tags as additional inputs for our baseline systems.

Multi-task Learning (MTL) Multi-task learning (MTL) is a widely discussed technique in the literature. Previous work (Changpinyo et al., 2018) shows that MTL can improve sequence tagging tasks in some cases. In our experiments, we try to jointly train a POS tagger and a semantic tagger which use a shared BiLSTM.

5 Experiments

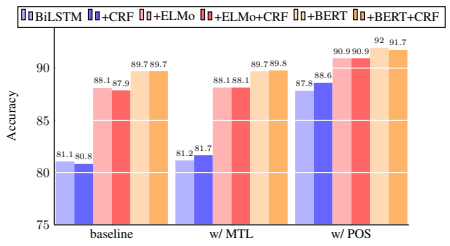
5.1 Experimental Setup

We conduct experiments on English and Chinese data separately. Since there are only about 2100 Chinese sentences and 1100 English sentences which are annotated, in order to achieve more stable tagging accuary for future comparison, we randomly split the whole dataset into 5 folds. One fold is a test set and the remaining serves as the training set where our model is trained on 85% instances and model selection is judged by the performance on the rest 15% instances. And then the tagging accuracy will be calculated using the best model on the selected fold. Finally, we report the average accuracy on these 5 folds.

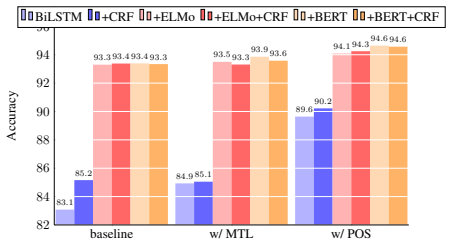
Built on the top of PyTorch (Paszke et al., 2017), we employ BiLSTM as our baseline model and all the models are trained for 8000 mini-batches, with a size of 32. Using the Adam optimizer (Kingma and Ba, 2015) and a cosine learning rate annealing method, we train the model with an initial learning rate chosen from $\{0.0001, 0.005, 0.001\}$. The details of parameters setting in different models are as follow: 1) the dimension of the hidden states of LSTM is set to 128 for each direction and the number of layers is set to 1; 2) the embeddings of POS tags are randomly initialized and has a dimension of 32 while the embeddings of words have a dimension of 300 and are initialized by the GloVe vectors[3] (Pennington et al., 2014) and pre-trained word vectors[4] (Li et al., 2018) for English and Chinese respectively[5]; 3) the parameters of BERT/ELMo are fixed during the training of our sequence tagging models; 4) for models with MTL, we directly optimize the sum of the losses for both POS tagging and universal semantic tagging.

5.2 Main Results

Figure 3 shows the overall performance of different models. Gold POS tags bring significant performance improvements, which is also verified by Huo and de Melo (2020). However, MTL can only slightly improve the overall results. When pre-trained contextualized word embeddings are utilized, the gap between different models becomes insignificant. Additionally, the significant improvement of English accuracy over previous state-of-the art is also attributed to the use of pre-training models: with the help of BERT, a simple BiLSTM tagger can be close to 92.0%-accurate for Chinese and 94.6% for English while without it, tagging accuracy of English data is around 85%.



(a) Chinese



(b) English

Figure 3: Averaged tagging accuracies.

[3] nlp.stanford.edu/projects/glove/

[4] github.com/Embedding/Chinese-Word-Vectors

[5] The embeddings missed in the pre-trained vectors are randomly initialized.

5.3 Comparative Analysis of Tagging Error

Empirical evaluation indicates competitive accuracy of our models. However, the result varies among different sem-tag categories and some of them remain at an extremely low level (Table 6).

To further improve the model's performance and have a better understanding of cross-lingual semantic representation, this section provides a fine-grained error analysis towards each underperforming sem-tag category.

Category of sem-tag	English	Chinese
ACT speech act	96.7%	86.8%
DXS deixis	64.9%	86.5%
ATT attribute	86.5%	82.6%
COM comparative	83.9%	88.2%
NAM named entity	91.9%	89.6%

Table 6: Tagging accuracies of five lowest sem-tag categories for English and Chinese

Properties of Chinese adjectives The low predication accuracy of ATT is largely attributable to the difficulties in differentiating IST and SST, especially in the light of high frequencies of adjectives in Chinese, which are a more complicated case compared to English adjectives. Usages of Chinese adjectives and corresponding sem-tags are shown in Table 7:

Usage	A	A+N	A+*de*+N
Narrow adjectives	IST	IST/SST	IST/SST
Distinct words	n.a.	IST	IST

Table 7: Usages and sem-tags of Chinese adjectives. "A" denotes adjective; "N" denotes noun; "de" is a Chinese particle denoting modification. In Mandarin Chinese, there are two sub-types of broad-sensed adjectives: narrow adjectives can both be used as predicates and modifiers while distinct words are only modifiers .

We propose practical strategies to improve the performance of our tagging model on differentiating IST and SST in Chinese. The first method is to establish a lexicon, based on the fact that whether an adjective can be used as a predicate is an inherent property. Thus it is possible to distinguish the use of IST and SST by simply referring to a lexicon. Another strategy is rule-based: an adnominal adjective is tagged SST only when it obtains a gradable reading. We stipulate the following rules: if tokens preceded by attribute adjectives are tagged INT, EQU, MOR and TOP, adjectives should be marked as SST. After uploading the lexicon and rules, the tagging accuracy of IST and SST raise from 68.8% and 63.1% to 81.4% and 77.9%. Overall accuracies after uploading adjective lexicon and rules are shown in Table 8.

	Baseline	w/MTL	w/POS
+BERT	90.2%	90.3%	92.7%
+BERT+CRT	90.0%	90.2%	92.7%

Table 8: Averaged Chinese tagging accuracies after uploading adjective lexicon and rules.

Named entity Table 9 shows the accuracy of each of NAM (named entity) for English and Chinese. Although named entities are regarded as one of the most frequently corresponding concepts shared by various languages (see §3), marked differences still exist:

- The accuracies of each sem-tag of English are generally higher than those of Chinese[6].
- English presents a lower diversity of performance (73.3%–98.0%) compared with Chinese (58.6%–97.9%).

Sem-tag	English	Chinese
PER person	98.00%	95.8%
GPE geo-political entity	92.1%	92.7%
GPO geo-political origin	88.0%	76.2%
GEO geographical location	73.3%	58.6%
ORG organization	94.3%	86.6%
ART artifact	76.1%	68.9%
HAP happening	n.a.	24.2%
UOM unit of measurement	93.2%	97.9%

Table 9: Accuracies of sem-tags under the NAM category for English and Chinese

We propose an explanation on why English and Chinese sem-taggers perform differently on NAM: named entities in English are identified by capitalization while Chinese not. Therefore, it is harder for Chinese to calculate the scope of proper names than English, and the overall accuracy is thus influenced. Moreover, it can also be inferred that Chinese is more sensitive to the length

[6] HAP is not included and will be discussed in the next paragraph.

of named entities given its difficulties in judging scope: sem-tags (PER, GPE and UOM) whose accuracies are higher than the average level, are commonly used to annotate one-token units while other below-average tags (GPO, GEO, ORG and ART) annotate multi-word proper nouns. On the contrary, English, with certain markers of named entites, shows that the decrease of accuracy with length is not as prominent as it of Chinese.

Sparse data input DXD of DXS, ITJ, HES and GRE of ACT, EQU of COM and HAP of NAM, whose presences are not enough for training and learning, need more diverse data as input in further research.

6 On Annotating Semantics

6.1 Helpfulness of Syntactic Features

The high-quality manual annotation and automatic tagging both indicate the importance of POS tags in the UST—the inter-annotator agreement and tagging accuracies increase after applying POS tags. Huo and de Melo (2020) believe this is because POS tags may facilitate semantic disambiguation though the extra syntactic information. However, what is not revealed is the underlying mechanism under which a syntactic feature can contribute to semantic analysis.

To investigate the impact of POS tags, 50 new sentences of WSJ and their Chinese counterparts are selected for a pilot study. Two annotators are asked to annotate them with or without the assistance of POS tags. Table 10 shows that POS tags have an impact on the inter-annotator agreements. This tendency is observed for both English and Chinese data.

Language	Type	IAA
English	+POS	96.1%
	−POS	95.2%
Chinese	+POS	93.6%
	−POS	90.8%

Table 10: The changes of inter-annotator agreements before and after the introduction of POS tags.

After a detailed investigation, we summarize the influences of POS tags on inter-annotator agreements as two points: (i) Some tokens have multi-dimensional semantic features and POS tags are likely to make annotators choose sem-tags related to POS features. For instance, *unable* may be annotated as NOT (negation) or POS (possibility). However, after the introduction of its POS tag, i.e. ADJ, two annotators are more likely to annotate it as IST, which is appropriate for most of adjectives, rather than NOT and POS; (ii) Gerunds which do not take arguments or are not modified by adverbs are more likely to bring challenges as it is difficult for annotators to determine whether event-related sem-tags or concept-related ones are more suitable for them. It is even more difficult for Chinese annotation in which verbs do not have inflected forms. All these can be easily solved by assigning POS tags.

In our view, the reason why POS contribute to semantic annotations can be traced to discussions of theoretical linguistics. Generally speaking, POS is category of words, whose identification has been a controversial problem for a long time in this area. Some linguists are in favor of a syntactic or distributional basis of POS (Harris, 1951; Edmonds, 1967) while others advocate a semantic or notional basis (Lyons, 1966). From a notion-based perspective, assigning forms to concepts, or POS tags and sem-tags to tokens, are all a process of categorizing and classifying objects referred by these tokens, which helps explain why POS tags have a significant influence on semantic sorts. In this regard, annotations are undoubtedly impacted by POS tags. Nonetheless, some researchers rebate it, believing that the notional definitions of POS are not applicable because of its unclearness. According to them, distribution, morphological features, grammatical functions are all useful criteria for the identification of POS. In our view, contradiction between notion-based and distribution-based approach leads to some difficulties in annotation. To avoid this, we applied POS tags which are automatically-generated by the Stanford CoreNLP tool (Manning et al., 2014b) to assist manual annotation.

However, though POS tags actually improve the inter-annotator agreement by regulating manual annotations of sem-tags in two ways, it is not clear whether they improve the quality of annotations—the first one increases the possibility of one option while the second one directly makes choices for annotators. To what extent more coarse-grained annotating standards contribute to annotations needs further research.

6.2 Challenges of Multilingual Annotations

Building comparative semantic representations across languages has been an important topic in recent years as a strategy to both contribute to semantic parsing and syntactic analysis. Existing approaches towards it can be roughly divided into three categories. First, crosslingual approach is proposed, which lends semantic annotation of a resource-rich language to an under-resourced language; see e.g. Damonte and Cohen (2018). However, crosslingual divergence between the lender and the borrower is likely to be retained to a considerable extent, especially for the languages which are phylogenetically distant. Another widely-discussed multilingual approach aims to achieve the goal by developing a comparable scheme of annotations for different languages, such as multilingual FrameNet (Baker et al., 1998) and multilingual WordNet (Miller, 1995), whose main limitation is that the semantic information represented is at the risk of oversimplifying since many in-depth properties are language-specific. The third one, the interlingual approach aims to find universal semantic frameworks for all languages. Yet it can be fairly difficult to find such appropriate interlingual frameworks.

In our view, these strategies are employed by researchers to study the major challenge i.e., the divergence of languages, encountered in representing multilingual data. And UST, which is in line with interlingual method, attempts to address it by a relatively shallow scheme. Despite the high inter-annotator agreements and tagging accuracies, there are still some divergences, which requires more in-depth study of multilingual annotation.

7 Related Work

UST is one of previous attempts of interlingua (Abzianidze and Bos, 2017), which is originally designed to provide necessary information for semantic parsing (Bjerva et al., 2016). Primary automatic sem-taggers are built using convolutional neural networks and deep residual networks (Bjerva et al., 2016). Later, in PMB project (Abzianidze et al., 2017), the authors propose a method of projecting automatically annotated semantic tags from a sentence to its sentence- and word-aligned counterparts. Following previous works, an updated universal semantic tagset is later proposed (Abzianidze and Bos, 2017), with a modification of deriving the tagset in a data-driven manner to disambiguate categories. In this work, a tri-gram based tagging model, TnT tagger (Brants, 2000), is also initially explored for bootstrapping utilization. In a recent study built on Bjerva et al. (2016), employing sem-tag in multi-task learning is found to be beneficial to both sem-tag task and other NLP tasks including Universal Dependency POS tagging, Universal Dependency parsing, and Natural Language Inference (Abdou et al., 2018). Overall, these studies indicate that sem-tags are effective in conducting various NLP tasks.

8 Conclusion

In this paper, we take Chinese into account to provide a more comprehensive tag set based on which we establish a reliable manually-annotated corpus, and show that promising performance of automatic semantic tagging is obtained after employing MTL as well as gold POS tag and leveraging pre-trained models. The overall success of this approach prompts a reflection of universality of different languages and operability of multilingual meaning representation: 1) UST is plausible in general partly because it is delexicalised and can thus represent phylogenetically languages after some adaptions; 2) universality is threatened to some extent because there are aligned but mismatched tokens between English and Chinese, which are caused by grammatical divergence, information loss of translation and different annotation strategies for MWE; and 3) innate crosslingual divergences still exist even in NAM's thought to be the most consistent pairs, which needs further exploration.

Though our work demonstrates the plausibility of developing a shared delexicalised and shallow annotation scheme to mitigate divergences across languages, it seems that more in-depth semantic analysis, especially lexicalised ones, may not be possible to be unified. We think a wider range of languages can be annotated after some minor adaptions of scheme. But it is still unknown how to get deeper processing information on this basis and thus develop an enhanced understanding of multilingual meaning representation.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments.

References

Mostafa Abdou, Artur Kulmizev, Vinit Ravishankar, Lasha Abzianidze, and Johan Bos. 2018. What can we learn from semantic tagging? In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4881–4889, Brussels, Belgium. Association for Computational Linguistics.

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In IWCS 2017 — 12th International Conference on Computational Semantics — Short papers.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638–1649.

Collin F. Baker and Michael Ellsworth. 2017. Graph methods for multilingual FrameNets. In Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, pages 45–50, Vancouver, Canada. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic tagging with deep residual networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3531–3541, Osaka, Japan. The COLING 2016 Organizing Committee.

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, Online. Association for Computational Linguistics.

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2642–2652, Melbourne, Australia. Association for Computational Linguistics.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Thorsten Brants. 2000. Tnt - A statistical part-of-speech tagger. CoRR, cs.CL/0003055.

Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: An empirical study. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2965–2977, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marta R. Costa-jussà, Cristina España-Bonet, Pascale Fung, and Noah A. Smith. 2020. Multilingual and interlingual semantic representations for natural language processing: A brief introduction. Computational Linguistics, 46(2):249–255.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jack Edmonds. 1967. Optimum branchings. J. Research of the National Bureau of Standards, 71B:233–240.

Daniel Flickinger, Yi Zhang, and Valia Kordoni. 2012. Deepbank: A dynamically annotated treebank of the wall street journal. In Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories, pages 85–96.

Zellig S Harris. 1951. Methods in structural linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9:1735–80.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.

Da Huo and Gerard de Melo. 2020. Inducing universal semantic tag vectors. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 3121–3127, Marseille, France. European Language Resources Association.

Nancy Ide and James Pustejovsky. 2017. Handbook of Linguistic Annotation, 1st edition. Springer Publishing Company, Incorporated.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Charles N Li and Sandra A Thompson. 1989. Mandarin Chinese: A functional reference grammar, volume 3. Univ of California Press.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 138–143. Association for Computational Linguistics.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo, and Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1520–1530, Lisbon, Portugal. Association for Computational Linguistics.

Meichun Liu. 2015. Tense and aspect in mandarin chinese. The Oxford Handbook of Chinese Linguistics, pages 274–289.

John Lyons. 1966. Towards a'notional'theory of the'parts of speech'. Journal of linguistics, 2(2):209–236.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014a. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014b. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pages 55–60.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the penn treebank. Computational Linguistics, 19(2):313–330.

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

Bruce Mitchell. 1985. Old English Syntax: Concord, the parts of speech, and the sentence. Oxford University Press, USA.

Tasnim Mohiuddin and Shafiq Joty. 2020. Unsupervised word translation with adversarial autoencoder. Computational Linguistics, 46(2):257–288.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 31:71–106.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Waltraud Paul. 2016. Where "complex" sentences are not complex and "subordinate". Coordination and subordination: Form and meaning—Selected papers from CSI Lisbon 2014, page 185.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the

Association for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. Computational Linguistics, 46(2):425–486.

Carlota S Smith and Mary S Erbaugh. 2005. Temporal interpretation in mandarin chinese. Linguistics, 43(4):713–756.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual bert transformation for zero-shot dependency parsing. pages 5725–5731.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. Linguistic Data Consortium, Philadelphia, PA.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The penn Chinese treebank: Phrase structure annotation of a large corpus. Natural Language Engineering, 11:207–238.

Appendix

In the supplemental material, we present the complete tailored universal semantic tag set for Mandarin Chinese (see Table 11).

ANA anaphoric

PRO	anaphoric & deictic pronouns: 他，她
DEF	definite: 这个，那人
HAS	possessive pronoun: 我弟弟，你学生
REF	reflexive & reciprocal pron: 自己，对方
EMP	emphasizing pronouns: 自己

ACT speech act

GRE	greeting & parting: 你好，再见
ITJ	interjections, exclamations:啊、哎呀
HES	hesitation: 额，*{.....}*
QUE	interrogative: 谁，什么，*{?}*

EVE events

EXS	untensed simple: 走、跑、休息
EXG	untensed progressive: 吃着、保持着
EXT	untensed perfect: 换了、见过

TNS tense & aspect

NOW	present tense: 现在
FUT	future tense: 将，将来
PRG	progressice: 在，着
PFT	perfect: 了、过

DSC discourse

COO	coordinate relations: *{，}*、*{；}*、所以
APP	appositional relations: *{——}*、*{，}*
BUT	contrast: 但是、然而

UNE unnamed entity

CON	concept: 狗、人
ROL	role: 学生、哥哥
GRP	group: 等、张三*{*、*}*李四和王五

DXS deixis

DXP	place deixis: 野外、沿着、前
DXT	temporal deixis: 过去、自从、后
DXD	discourse deixis: 首先、其次

LOG logical

ALT	alternative & repetitions: 另、再
XCL	exclusive: 只、仅仅
NIL	empty semantics: *{。}*、*{《》}*
DIS	disjunction & exist. quantif.: 或、某
IMP	implication: 如果、除非、当
AND	conjunction & univ. quantif.: 并且、所有

MOD modality

NOT	negation: 不、没有
NEC	necessity: 得、该
POS	possibility: 能、可能、应该

ATT attribute

QUC	concrete quantity: 二，六百万
QUV	vague quantity: 一些，几
COL	color: 红，浅蓝
IST	intersective: 公，大型
SST	subsective: 高，热，可笑
PRI	privative: 假，前，副
DEG	degree: *2*米高，*8*个月大
INT	intensifier: 非常、很
SCO	score: *3-0*、*100*分

COM comparative

EQU	equative: 这么、这样、和他一样高
MOR	comparative positive: 更
TOP	superlative: 最
ORD	ordinal: 第一、首次

NAM named entity

PER	person: 包拯、狄仁杰
GPE	geo-political entity: 北京、日本
GPO	geo-political origin: 华裔、东乡族
GEO	geographical location: 长江、尼罗河
ORG	organization: 宜家、欧盟
ART	artifact: *ios 7*、安卓
HAP	happening: *2017*青歌赛
UOM	unit of measurement: 米、个、美元
CTC	contact information: *110*、*info@mail.com*
URL	*URL: http://pmb.let.rug.nl*
LIT	literal use of names: 他的名字是张三
NTH	other names: 图*(1)*

TIM temporal entity

DAT	full date: *2019*年*4*月*11*日、*11/04/19*
DOM	day of month: *12*月*27*日
YOC	year of century: *2019*、*2019*年
DOW	day of week: 星期四、周四
MOY	month of year: 四月
DEC	decade: *90*年代
CLO	clocktime: 十点、*8:45*

ADD additional

MAN	manner: 按照、根据、本着
RES	reason: 因、因为、由于
AIM	aim: 为、为了、为着
OBJ	object : 对、和、跟、替
COM	comparation: 比、较
MOD	modification: 的、地、得

Table 11: Modified tag set for Chinese