Negation Scope Resolution for Chinese as a Second Language

Mengyu Zhang[1], **Weiqi Wang**[2], **Shuqiao Sun**[3] and **Weiwei Sun**[4*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Chinese as a Second Language, Peking University
[3]Department of Chinese, Bukkyo University
[4]Department of Computer Science and Technology, University of Cambridge
zhang.mengyu@pku.edu.cn; ws390@cam.ac.uk

Abstract

This paper studies Negation Scope Resolution (NSR) for Chinese as a Second Language (CSL), which shows many unique characteristics that distinguish itself from "standard" Chinese. We annotate a new moderate-sized corpus that covers two background L1 languages, viz. English and Japanese[1]. We build a neural NSR system, which achieves a new state-of-the-art accuracy on English benchmark data. We leverage this system to gauge how successful NSR for CSL can be. Different native language backgrounds of language learners result in unequal cross-lingual transfer, which has a significant impact on processing second language data. In particular, manual annotation, empirical evaluation and error analysis indicate two non-obvious facts: 1) L2-Chinese, L1-Japanese data are more difficult to analyze and thus annotate than L2-Chinese, L1-English data; 2) computational models trained on L2-Chinese, L1-Japanese data perform better than models trained on L2-Chinese, L1-English data.

1 Introduction

In an increasingly globalized world, non-native speakers and writers all over the world produce a huge amount of second language (L2) data every day. There is naturally a need to automatically annotate such large-scale *atypical* data with rich lexical, syntactic, semantic and even pragmatic information. High-performance automatic annotation of such data, from an engineering perspective, enables deriving high-quality information by structuring this specific type of data, and from a scientific perspective, enables quantitative studies for Second Language Acquisition (SLA), which is complementary to hands-on experiences in interpreting second language phenomena (Gass, 2013).

It is insufficient to directly apply models designed for first languages to handle the second language data because a second language is a standalone linguistic system (Selinker, 1972) that distinguishes itself from both the source and target languages in linguistic features. We need second language-specific methodologies by taking its characteristics into account. This direction has been recently explored, including syntactic and semantic parsing for English as a Second Language (Nagata and Sakaguchi, 2016; Berzak et al., 2016a; Zhao et al., 2020) and semantic role labeling for Chinese as a Second Language (hereafter L2-Chinese or L2-CHI for short) (Lin et al., 2018).

Negation is a ubiquitous phenomenon in spoken text (27.6 per 1000 words) and in written text (12.8 per 1000 words) (Tottie, 1991). As a type of extra-propositional semantics, negation is relevant to analyze factuality of propositions and thus relevant to accurately understand natural languages which has been proven useful for several NLP applications such as sentiment analysis (Wiegand et al., 2010), machine reading comprehension (Morante and Daelemans, 2012a), automatic question and answer (Councill et al., 2010), etc. In the NLP community, the morpheme or word that expresses negation is called the negation cue, which can be a single word or multiple words. A negation is related to an event, i.e. so-called negation event, which is identified by some keys words, like *need* in (1). If we want to know the complete information conveyed by a negation event such as *what* is not *needed* in (1), we need to identify a **scope** in the sentence, which is a set of words. To be more precise, the negation scope is the maximum part(s) of the sentence that are influenced or negated by negation cue. In (1), the negation scopes are marked with square brackets.

*Work done at Peking University.
[1]Data is available at https://github.com/pkucoli.

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1–10

(1) [We needs] actions and not [thoughts].

In this paper, we study Negation Scope Resolution (NSR) for L2-Chinese, the goal of which is to automatically identify the scope of a negation cue. To the best of our knowledge, this is the first study on NSR for a second language. We build a new moderate-sized Chinese corpus that contains manual negation annotations for 8000 sentences written by non-native speakers, as well as their corrections by native speakers. A high inter-annotator agreement is achieved, suggesting the robustness of language comprehension of CSL. All annotated sentences are standard Chinese or L2-Chinese. We consider two typologically distant languages, i.e. English and Japanese, as background native languages of language learners. We use L1-English (L1-ENG for short) and L1-Japanese (L1-JPN for short) to denote them respectively[2]. We also consider the corrected sentences by Chinese native speakers and use "$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$[3], L1-ENG" and "$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$, L1-JPN" to denote them. The following are an example of L2-CHI, L1-JPN (2a) as well as its correction (2b).

(2) a. 换 言 说 , 没有 [宗教
Change text speak , have-not religion
生活 与 日常 生活 差距]。
life and ordinary life difference.
'In other words, there is no difference between religious life and ordinary life'

b. 换 言 说 , [宗教 生活 与
Change text speak , religion life and
日常 生活 之间] 没有
ordinary life between have-not
[距离]。
difference.
'In other words, there is no difference between religious life and ordinary life'

With the availability of high-quality annotations, we study neural NSR models which have achieved significant advances during the past several years. We build a state-of-the-art system that is based on BERT (Devlin et al., 2018), and study NSR for first and second languages across a wide range of setups. Evaluation gauges how successful NSR for Chinese can be by applying state-of-the-art neural techniques. In particular, we find that the L1 background has a significant impact on the L2 outputs: Models trained on L2-CHI, L1-JPN data achieve better performance than L2-CHI, L1-ENG data. An error analysis indicates that some interesting cross-lingual transfer phenomena resulted from the difference between Chinese and English or Japanese play a significant role. Further linguistically-informed analyses suggest several directions for future study on NLP for second languages.

2 Related Work

2.1 Negation Scope Resolution

There are three corpora annotated with negation cues and negation scopes for English (Morante and Daelemans, 2012b; Konstantinova et al.; Vincze et al., 2008) and one for Chinese (Zou et al., 2015). All of them focus on first languages only. Tab. 1 summarizes the sizes of existing corpora as well as our new creation.

Previous approaches to automatic NSR can be categorized into word-based, syntax-based and semantics-based approaches. Similar to many word-based solutions (Tan et al., 2018; He et al., 2017) to Semantic Role Labeling (Carreras and Màrquez, 2004), NSR can be cast as a sequence labeling problem over word sequence. Each word is assigned a position label, e.g. BEGIN-SCOPE, which indicates if the word s in the scope of a particular cue. Representative sequence labeling models, e.g. linear-chain CRF, semi-Markov CRF and latent variable CRF, have been evaluated (Li and Lu, 2018). Neural models have also been explored in (Fancellu et al., 2016, 2017): for each token the corresponding lemma PoS and cue features are fed into BiLSTM to do binary classification to indicate whether the token is in the negation scopes.

Syntactic parsing provides valuable structural information for semantic analysis. It is a usual case that a negation scope is correlated to a single constituent. Following this property, NSR can be treated as a discriminative ranking problem over constituents returned by a phrase-structure parser. Read et al. (2012) used some hand-written heuristic features derived from constituent syntactic trees to guide a statistical ranker (Read et al., 2012). Dependency-based analysis can also provide effective syntactic analysis. For example, Lapponi et al. (2012) augmented a word-based sequence model with dependency tree based features (Lapponi et al.,

[2]L1 is short for *first language*.

[3]Note that $\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$ is different from L2-Chinese because it is corrected by native speakers and also different from standard L1-Chinese to some extent because the source usage, e.g. lexical selection, has a significant impact.

	Language	#Sentence	#Negation	Source
CD-SCO	English	4,423	1,160	Conan Doyle stories
SFU Review	English	17,263	3,527	Consumer product reviews
BioScope	English	20,924	3,114	Medical & biological texts
CNeSp	Chinese	16,841	4,517	Product, scientific & financial texts
Ours	L2-CHI, L1-ENG	2,098	2,253	
	$^{\text{CHI}}\text{L2}\Rightarrow\text{L1}$, L1-ENG	2,098	2,253	
	L2-CHI, L1-JPN	1,888	2,181	
	$^{\text{CHI}}\text{L2}\Rightarrow\text{L1}$, L1-JPN	1,888	2,181	
	L2/L1-Chinese (Total)	7,972	**8,868**	Learners' essay

Table 1: A comparison of sizes of existing and our corpus. "#Sentence" shows the total numbers of sentences; "#Negation" shows the total numbers of negation expressions (=cues).

2012) .

Though the scope of negation is a type of important semantics, it is not included in almost all sentence-level SemBanks, including English Resource Semantics (Flickinger et al.), Groningen Meaning Bank (Bos et al., 2017) and Abstract Meaning Representations (Burns et al., 2016), due to either theoretical or practical considerations. Nevertheless, the output structures provided by a semantic parser have been shown very powerful to assist the discovery of negation scopes (Packard et al., 2014; Basile et al.).

2.2 NLP for Second Languages

Despite the importance of second languages at both the scientific and engineering levels as mentioned earlier, it is little systematically studied in the NLP community. Second language, the language system developed by a learner of a second language, preserves linguistic characteristics from both the native and target languages and thus it is a unique linguistic organization (Selinker, 1972). Therefore, the automatic processing of learner texts is not as simple as directly utilizing the existing machinery designed for native languages (VanPatten and Jegerski, 2010). Before any further exploration, the task should be well formulated based on sound theoretical analysis. There are several attempts to set up annotation schemes for different linguistic layers for learner languages, such as PoS tags and syntactic information (Rosen et al., 2014). Syntactic and semantic parsing for English as a Second Language (Nagata and Sakaguchi, 2016; Berzak et al., 2016b; Zhao et al., 2020) and semantic role labeling for CSL (Lin et al., 2018) have been explored based on annotated corpora.

3 The New Corpus

3.1 The Raw Data

Lang-8 (https://lang-8.com/) is a *language-exchange* social networking platform, where second language learners of different languages can put their essays and receive modifications from native speakers. There are about 68,500 Chinese learners registering on this platform, most of whose mother tongues are English (ca. 15,500 users) and Japanese (ca. 25,700 users). Following (Mizumoto et al., 2011; Lin et al., 2018), we build a large-scale parallel Chinese learners' corpus by extracting and cleaning the original sentences (L2) written by Mandarin learners and the corresponding revised version (L1) from native speakers. It consists of 717,241 sentence pairs from writers of 61 different native languages, of which English and Japanese constitute the majority. We carefully select 4,000 sentence pairs and manually annotate them with negative cues and scopes based on parallel linguistic analysis. We notice that learners' native languages can have a great impact on grammatical errors and hence automatic prediction of negative cues and scopes. Therefore, our corpus includes two typologically distinct languages, English and Japanese, each of which has a sub-corpus consisting of 2000 sentences.

3.2 The Annotation Process

Two graduate students are employed to annotate the selected sentences: One annotator has rich experience in computational research on negation, while the other specializes in linguistics. We defined the annotation guideline mainly by combining experiences from the CNeSp Corpus (Zou et al., 2015) and the CD-SCO Corpus (Morante et al., 2011) and

adding some modifications based on the unique characteristics of the parallel data. Initially, the two annotators separately annotate the same 800 sentences: 200 sentences are "L2-CHI, L1-ENG" and "$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$, L1-ENG" pairs, 200 are "L2-CHI, L1-JPN" and "$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$, L1-JPN" pairs. After this round, annotators compared their annotations and discussed incompatible phenomena to produce the initial annotation guideline. Then the two annotators proceeded to annotate an 800-sentence set independently based on this guideline. It is on this set that we calculate the inter-annotator agreement, as shown in Tab. 2. Based on the disagreement of the second annotation round, we made some minor adjustments to the guideline. Since the agreement is relatively satisfactory, the rest sentences only include single blind annotations. Noted that the annotation was based on the segmentation results produced by the Stanford CoreNLP tool (Manning et al., 2014), and during the annotation process wrong segmentation was corrected.

3.3 Annotation Guidelines

Our annotation guidelines are inspired by the CNeSp Corpus (Zou et al., 2015) and the CD-SCO Corpus (Morante et al., 2011) with exceptions considering characteristics of CSL and Mandarin Chinese. In the following, we will use double underline to mark the negation cue, and square brackets to mark the negation scope.

3.3.1 Annotation of Negation Cues

The words that express negation are negation cues. We do not consider word-internal morphemes that express negation-related meanings. For example, although the negation-like morpheme "无" is included in "无价/*priceless*" and "无话可说/*have nothing to say*", its contribution to the meanings of the word as a whole is non-compositional. Therefore, we do not consider it as a cue. When we meet sentences containing such fixed expressions, we never mark negative cues within a single word. Besides, there are some words that can be mixed up with negative expressions, e.g., "杜绝/*eliminate*", which are not included in our research.

3.3.2 Annotation of Negation Scopes

We try to cover the longest relevant scope in order to capture the exact meaning of a negation. When there are more than one negation cues in one sentence, we will annotate them separately and focus on the most relevant scope of each cue. For example, "没有不必要的东西/*no unnecessary things*", the scope of the second cue is "必要/*necessary*". If the negated verb is the main verb of the sentence, the entire sentence is under the scope. In addition, the scope can be discontinuous, which means that if the subject or the object are elliptical but explicit in another conjunct clause, then we will mark it over clause, as shown in (3). If the sentence contains different clauses and only one of them contains a cue, then we will mark the most relevant part, as shown in (4). But when sub-clauses contain words that are related to the negated fact, such as *save* and *except*, they are included in the negation scope to cover the largest negation scope. In Chinese, the position of some adverbials are very flexible, thereby making it difficult to tell their smallest and specific semantic scopes. Therefore, if the sentence contains different clauses and sentence-level adverbials, we will assign them to their closest clause, as shown in (4). When annotating the negation scope of the negation cue in the second clause, the at the beginning of the sentence is excluded. Conjunctions, e.g., adversative conjunction (虽然/*although*) and cause conjunction (因为/*because*), are excluded from the scope. As for imperative and interrogative sentences, the verb (e.g., 请/*please*) and interrogative pronouns (e.g., 为什么/*why*) are included in the negation scope.

(3) [我] 一直 呆 在 家里 , 没 [做
I all-along stay at home , not do
坏事 啊]。
troubles A.
'I stayed at home all along and didn't make any troubles.'

(4) 在 日本 , 年轻人 对 高档 商品
In Japan, young-people for luxuries nearly
没有 兴趣 , [他们] 没有 [购买 欲]。
not interest , they not buy desire.
'In Japan, young people nearly have no interest in luxuries. They have no desire to buy.'

3.4 Inter-Annotator Agreement

We measured the inter-annotator agreement (IAA) of negation cues and scopes between two independent annotators using two measurements: (1) Kappa (Cohen, 1960) and (2) precision, recall as well as F_1 of one annotator treating the second one as golden standard, which is used as the evaluation metric for evaluating the automatic NSR system (see §5.1). Tab. 2 is a summary of the measurement results, reflecting a high agreement. Note that

the length of sentences in our corpus is shorter because of limited language capability of L2 learners. The ratio of disagreement in sentences written by Japanese native speakers is higher than that of sentences written by English native speakers, which may result from the varying complexity of expressions between them.

3.5 Disagreement Analysis

The disagreement between two annotators all occurs in the scope, which mainly appears in the recognition of elliptical subject or object and adverbial adjunct. Ellipsis without any formal markers often occurs in Chinese, especially when there is more than one sub-clauses sharing the same subject, causing the annotators to recognize the subject and object differently. In Ex. (5) from an English native speaker, "很多同事/*many colleagues*" is omitted in the last clause where the scope of "不/*no*" exists.

(5) 很多 同事们 没有 注册， 而且 到
Many colleagues not register, and arrive
中国 晚 了， 不 [会 登录 谷歌 邮箱]。
China late LE, not can sign google email.
'Many colleagues haven't registered and arrived in China late. They couldn't sign in g-mail.'

Another type of disagreement results from the multiple interpretations of sentence meanings. For example, in sentence (6), written by a Japanese native speaker, "有的/*something*", as a pronominal part, refers to some of the "暑假的美好事情/*wonderful memories of summer holidays*", which can also be extended into "有的暑假的美好回忆/*some of the wonderful memories of summer holidays*". This makes annotators have different judgments of the scope.

(6) 暑假 的 美好 的 事情，
Summer-holidays De wonderful De things,
[有的 已经 想] 不 [起来]。
some already remember not QILAI.
'I have not remembered some of the wonderful things in the summer holiday.'

Disagreement can also result from the form regulations, like the Ex. (7). In some southern dialects of China, "有/*have*" can be added before a verb, which is not allowed in Mandarin Chinese. However, in the construction like "有没有/*have or have not + verb*", "有/*have*" cannot be omitted. Two annotators thus interpreted the negation scope differently.

(7) 我 不 知道 [大家 有] 没有 [了解 在
I not know us have not know in
国内 给 外国人 的 实习
domestic give foreigners DE intern
机会]。
opportunities.
'I don't know whether they know opportunities for foreign interns.'

Our corpus contains sentences written by learners who have not reached native-like proficiency, and therefore there are some non-native expressions that are difficult to disambiguate, contributing part of the annotation disagreements. In Ex. (8), written by a Japanese native speaker, both "没有/*do not have*" and "是/*is*" can be treated as the main verb of the sentence. If "没有/*do not have*" is the main verb, then the sentence can be understood as "nothing can be resistant to the antibiotic's effect". In another condition, "抗生素的效果几乎没有/*antibiotic does not affect*" is the adjunct of "大敌/*enemy*".

(8) [这 是 抗生素 的 效果 几乎] 没有 [的
This is antibiotic DE effect nearly no DE
大敌]。
enemy.
'The antibiotic's effect is obvious and it has no enemies.'

4 The Neural NSR Model

In this paper, we focus on the state-of-the-art solution that leverages on neural word-based tagging to solve the problem. The overview of our neural model is shown in Fig. 1. Formally, our task is to predict a sequence y given a word-PoS-cue pair $\langle w, p, c\rangle$ as input. Each $y_i \in y$ is a binary label to show whether the word is inside the negation scope. A sentence may include more than one negation cues. To make sure the neural model is aware of which negation cue is in question, the value of $c_i \in c$ is set to 1 if the current cue is in question, and 0 otherwise. As Fig. 1 shows, we use BERT (Devlin et al., 2018) or ELMo (Peters et al., 2018) or randomly initialized feedforward layers for word embedding of w_i, randomly initialized feedforward layers for PoS embedding of p_i, randomly initialized feedforward layers for cue embedding of c_i. Then we use feed forward layers to compress word embedding vectors of BERT and ELMo to a lower dimension. The concatenation of feature embeddings is fed to an encoder (e.g., BiLSTM). The encoder's outputs then go through feedforward lay-

	C. F_1	S. F_1	T. F_1	C. Kappa	S. Kappa
CD-SCO (Morante and Daelemans, 2012b)	94.88	85.04	91.53	-	-
SFU Review (Konstantinova et al.)	92.79	81.88	-	92.70	87.20
BioScope (Vincze et al., 2008)	98.65	95.91	-	-	-
CNeSp (Zou et al., 2015)	-	-	-	95.00	93.00
L2-CHI, L1-ENG	100.00	92.55	97.12	100.00	96.13
$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$, L1-ENG	100.00	92.55	97.71	100.00	95.65
L2-CHI, L1-JPN	100.00	90.09	94.62	100.00	92.15
$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$, L1-JPN	100.00	90.09	94.35	100.00	92.32

Table 2: Inter-annotator agreement with respect to negation cue and scope. Previous negation corpora reported cue-level F_1 (C. F_1) at 91%-95%, scope-level F_1 (S. F_1) at 76%-85%, token-level F_1 (T. F_1) at 88%-92%, and kappa at 87%-91%.

ers for classification. The cross-entropy loss is utilized for training the whole model.

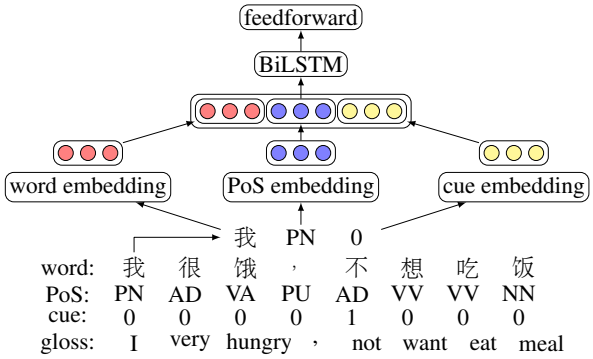

Figure 1: The architecture of our neural NSR model. The meaning of the Chinese sentence is *I'm very hungry but I don't want to have a meal.*

5 Experiments

5.1 Setup

The evaluation metric of the auto NSR system is the same as the *SEM 2012 shared task (Morante and Daelemans, 2012b). We will reproduce the sentence according to the number of negation cues in it and predict each one separately. The system is evaluated with respect to precision, recall, and F1 at both scope-level and token-level. For a scope to be correctly classified, all tokens in a sentence must be correct. Significantly, if an empty negation scope is predicted incorrectly, it counts as precision otherwise, it counts as recall. In our Chinese corpora there are no empty negation scopes, and thus precision is 100%. Therefore, we report the scope recall instead of scope F_1. A sentence may include more than one negation cues. Then we will create as many copies as the number of negation cues.

The word segmentation is manually annotated as provided by our corpus. The PoS tags are produced by the Stanford CoreNLP toolkit (Manning et al., 2014) of which the macro-averaged F_1 over all tree-banks is 90.99% on UPoS. We use BERT as the pre-training models and set max sequence length to 128. BERT is fine-tuned during training the NSR models. We also utilize ELMo and re-train it using the Chinese Gigaword corpus (Huang, 2009). We use 2 feed forward layers to transfer the large output dimension to a lower dimension that is 768-256-50 for BERT and 1024-256-50 for ELMo. The dimension of PoS tag embedding and cue embedding is 50 too. Then the concatenation of word-PoS-cue embedding with dimension of 150 goes through 1 layer BiLSTM with 256 dimensional hidden units, 1 feed forward layer and a softmax layer for predicting the binary output distribution. The dropout of feed forward layers and BiLSTM is 0.2. The neural network is implemented using PyTorch. During training, gradients are updated using Adam (Kingma and Ba, 2014). The learning rate of BERT is 0.00001. The Adam with an initial learning rate of 0.001.

5.2 Main Results on English Data

To evaluate the effectiveness of our model, we conduct experiments on the quite standard English CD-SCO data set (Morante and Daelemans, 2012b). Read et al. (2012), Li and Lu (2018) and we use the TnT tagger (Brants, 2000) and Fancellu et al. (2016) uses the GENIA tagger (Tsuruoka et al., 2005) to produce PoS tags. The results are summarized in Tab. 3. We can clearly see the effectiveness of our architecture.

Model	S-F_1	T-F_1
Read et al. (2012)	72.2	85.3
Fancellu et al. (2016)	77.8	88.7
Packard et al. (2014)	78.7	88.2
Li and Lu (2018)	82.0	88.6
Khandelwal and Sawant (2020)	- -	92.4
Random+BiLSTM	76.8	89.4
ELMo+BiLSTM	82.6	91.6
BERT+BiLSTM	83.7	92.5

Table 3: NSR results on the CD-SCO data. "S-F_1" is the f-score of scope prediction; "T-F_1" is the f-score of token classification.

5.3 Main Results on the Chinese Data

For all experiments on Chinese, we use 5-fold cross-validation. At each fold, as Fig. 2 shows, we build three training scenarios: (1) using all four subsets, (2) $\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$ or L2 subsets, (3) "L2-CHI, L1-ENG" or "$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$, L1-ENG" or "L2-CHI, L1-JPN" or "$\text{CHI}_{\text{L2}\Rightarrow\text{L1}}$, L1-JPN" subset.

Though another existing negation corpus, i.e. CNeSp (Zou et al., 2015), is available, we do not use it for either training nor test. For a significant number of linguistic phenomena, we have different analyses from CNeSp. Therefore the integration of the two corpora is not straightforward.

We conduct experiments according to the three training scenarios demonstrated by Fig. 2. The BERT-BiLSTM model is utilized. Tab. 4 and 5 show the overall results. As shown in Tab. 4, the token-level F1 score obtained on our corpus is 93.6% which is slightly better than that on the CD-SCO data (92.5%). This may be because that more sentences are available for training (3544 of our corpus, while 982 of the CD-SCO corpus) and shorter sentences are utilized for evaluation (18.0 words of our corpus, while 22.5 words of the CD-SCO corpus). Another important factor is the complexity of second languages. The outputs of second language learners are usually simpler than the ones of native speakers.

We assume that observations from empirical evaluations represent the overall characteristics of second language and native language based on the following facts:

- All sentences in our corpus are randomly selected from the Lang-8 website that contains large-scale sentences and therefore there is no sampling bias in constructing English parallel subset and Japanese parallel subset.
- All the setups and hyper-parameters of each run are the same and therefore there is no modeling bias.
- We use 5-fold cross-validation to prevent train-test split bias.
- We experiment scenarios on several models like BERT-BiLSTM, ELMo-BiLSTM, etc. The same impact is observed on different models.

5.4 An Analysis of Cross-Lingual Influence

Results in Tab. 5 show a distinct and unique phenomenon that models trained on L1-JPN data can predict L1-ENG relatively well, while models trained L1-ENG data perform rather badly on L1-JPN data. This is consistent among all neural models, including BERT-BiLSTM, ELMo-BiLSTM and Random-BiLSTM.

We examine L1-JPN sentences that are wrongly predicted by models trained on L1-ENG data while correctly handled by models trained on L1-JPN data, in order to explain why there is drop on the result in boldface in Tab. 5. Possible reasons for these errors can be a result of the cross-lingual transfer from Japanese to Chinese, which furnishes sentences with characteristics that are not shared by the English language. In the following examples, we use square bracket to mark the golden scope annotation and underline to mark the wrong prediction of trainset-ENG.

(9) 所以 大多数 的 人 觉得 [受] 不
So most DE people believe stand not
[了 日本 的 夏天]。
LE Japan DE summer.
'So most people believe that they cannot stand the summer of Japan.'

(10) [我 还] 没有 [上 小学]
I have not attended elementary-school
的 时候。
DE time.
'When I have not attended elementary school.'

(11) [我们] 不 [应该 恐怕 说错 还有
We not should afraid speak-wrong and
不好意思 的 事]。
embarrassed DE thing.

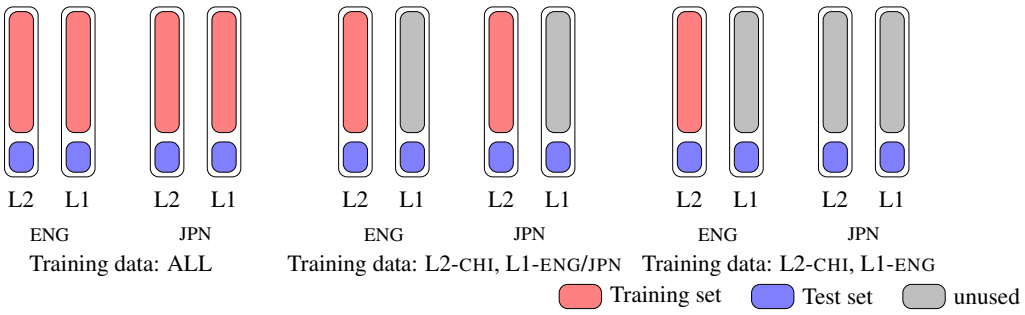

Figure 2: Three different training scenarios.

Train \ Test	L2-CHI, L1-ENG	CHI$_{L2\Rightarrow L1}$, L1-ENG	L2-CHI, L1-JPN	CHI$_{L2\Rightarrow L1}$, L1-JPN
ALL	75.8/93.2	76.1/93.4	75.3/93.6	74.7/93.6
L2-CHI, L1-ENG/JPN	75.2/92.9	75.7/93.3	74.8/93.3	74.4/93.6
CHI$_{L2\Rightarrow L1}$, L1-ENG/JPN	74.8/93.0	75.9/93.4	75.1/93.2	74.5/93.4

Table 4: The scope recall/token F_1 scores obtained by the BERT-BiLSTM.

Train \ Test	L2-CHI, L1-ENG	CHI$_{L2\Rightarrow L1}$, L1-ENG	L2-CHI, L1-JPN	CHI$_{L2\Rightarrow L1}$, L1-JPN
L2-CHI, L1-ENG	73.4/67.8/61.3	73.7/69.4/62.9	**71.6/64.9/56.3**	**71.1/64.5/56.9**
CHI$_{L2\Rightarrow L1}$, L1-ENG	73.4/68.1/61.6	73.8/69.7/62.4	**70.7/65.1/58.2**	**71.0/65.4/57.6**
L2-CHI, L1-JPN	73.2/65.6/57.3	74.7/68.0/60.7	75.2/68.5/62.8	74.2/68.4/61.5
CHI$_{L2\Rightarrow L1}$, L1-JPN	72.9/65.1/58.0	74.0/68.3/60.8	74.6/68.4/59.8	74.1/68.5/60.7

Table 5: The scope recall scores obtained by the BERT-BiLSTM/ELMo-BiLSTM/Random-BiLSTM.

'We should not be afraid of making oral mistakes or being embarrassment.'

Errors can be categorized into three types: redundant subject, the incorrect judgment of the relative clause, and neglect of parallel components. For example, omitting subjects is common in Japanese while unusual in English, so the system tends to find a subject when testing sentences (see (9)). Besides, the verb phase attributes always come before the core component in Japanese and Chinese, which will be expressed in "DE/的" construction in Chinese and "NO/の" in Japanese, while this phenomenon is replaced by relative clause in English (see (10)). The system that has little experience in this structure thus tends to have difficulties detecting the exact scope. Additionally, the parallel descriptive components connected by coordinating conjunctions in Japanese usually share one core component that is "的事/the thing" which doesn't exist in Chinese and English (see (11)). Therefore, the system may stop when finding the first component that is a complete verb phrase structure.

Based on observations of empirical and error analysis, we give an initial explanation that (1) the cross-lingual transfer occurs unequally to learners with different native backgrounds and on our work there is more cross-linguistic transfer related to negation on Japanese-Chinese than English-Chinese, and (2) the cross-lingual transfer phenomenon on the second languages will directly affect the performance of the computational system which indicates second languages are a stand-along linguistic systems and it distinguishes themselves from both the source and target languages. In summary, when we study computational models for processing second languages it is essential to take the native background into consideration.

6 Conclusion

In this paper, we construct a Chinese negation corpus with sentences written by English, Japanese and Chinese native speakers respectively. Compared to previous negation corpus, our corpus has several highlights: (1) our corpus has the largest scale and higher inter-annotator agreement; (2) our corpus is the first corpus that focuses on second lan-

guage and contributes parallel sentences. Also, we present a state of the art system for automatic NSR on the English benchmark data. We evaluate our system on our corpus on cross-language training scenarios where a model is trained and tested on different sub-corpora. Results and error analysis show that cross-lingual transfer has a significant impact, but not on all languages equally. The observation that different first language background will have a different impact on the second language acquisition will give inspirations on NLP for second languages.

Acknowledgement

We thank the anonymous reviewers for their useful feedback and suggestions.

References

Valerio Basile, Johan Bos, Kilian Evang, and Noortje. Ugroningen: Negation detection with discourse representation structures. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics, pages 301–309.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016a. Universal dependencies for learner english. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1), pages 737–746.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016b. Universal dependencies for learner english. arXiv:1605.04278.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. 2017. The Groningen Meaning Bank. Springer Netherlands.

Thorsten Brants. 2000. Tnt: A statistical part-of-speech tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing, page 224–231.

Gully A. Burns, Ulf Hermjakob, and José Luis Ambite. 2016. Abstract meaning representations as linked data. In International Semantic Web Conference.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In Proceedings of the Eighth Conference on Computational Natural Language Learning, pages 89–97.

J Cohen. 1960. Kappa: Coefficient of concordance. Educ Psych Measurement, 20:37–46.

Isaac G Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In Proceedings of the workshop on negation and speculation in natural language processing, pages 51–59.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1), pages 495–504.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 58–63.

Dan Flickinger, Emily M. Bender, and Woodley Packard. English resource semantics. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts, pages 1–5.

Susan M Gass. 2013. Second language acquisition: An introductory course. Routledge.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1), pages 473–483.

Chu-Ren Huang. 2009. Tagged chinese gigaword corpus 2.0.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 5739–5748, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. A review corpus annotated for negation, speculation and their scope. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), pages 3190–3195.

Emanuele Lapponi, Erik Velldal, Lilja Øvrelid, and Jonathon Read. 2012. UiO 2: Sequence-labeling negation using dependency features. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics (SemEval 2012), pages 319–327.

Hao Li and Wei Lu. 2018. Learning with structured representations for negation scope extraction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 533–539.

Zi Lin, Yuguang Duan, Yuanyuan Zhao, Weiwei Sun, and Xiaojun Wan. 2018. Semantic role labeling for learner chinese: the importance of syntactic parsing and l2-l1 parallel data. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3793–3802.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Association for Computational Linguistics (ACL) System Demonstrations, pages 55–60.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated japanese error correction of second language learners. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 147–155.

Roser Morante and Walter Daelemans. 2012a. Annotating modality and negation for a machine reading evaluation. In CLEF (Online Working Notes/Labs/Workshop).

Roser Morante and Walter Daelemans. 2012b. Conandoyle-neg: Annotation of negation in conan doyle stories. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul.

Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope: Guidelines v1. Computational linguistics and psycholinguistics technical report series.

Ryo Nagata and Keisuke Sakaguchi. 2016. Phrase structure annotation and parsing for learner english. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1), volume 1, pages 1837–1847.

Woodley Packard, Emily M. Bender, Jonathon Read, Stephan Oepen, and Rebecca Dridan. 2014. Simple negation scope resolution through deep parsing: A semantic solution to a semantic problem. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1), pages 69–78.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. CoRR.

Jonathon Read, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. UiO1: Constituent-based discriminative ranking for negation resolution. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics (SemEval 2012), pages 310–318.

Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. Language Resources and Evaluation, 48(1):65–92.

Larry Selinker. 1972. Interlanguage. IRAL-International Review of Applied Linguistics in Language Teaching, 10(1-4):209–232.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In Thirty-Second AAAI Conference on Artificial Intelligence.

Gunnel Tottie. 1991. Negation in English speech and writing: A study in variation, volume 4.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin Dong Kim, Tomoko Ohta, and Jun Ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In Advances in Informatics, 10th Panhellenic Conference on Informatics.

Bill VanPatten and Jill Jegerski. 2010. Research in second language processing and parsing, volume 53.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics, 9:S9–S9.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In Proceedings of the workshop on negation and speculation in natural language processing, pages 60–68.

Yuanyuan Zhao, Weiwei Sun, Junjie Cao, and Xiaojun Wan. 2020. Semantic parsing for English as a second language. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6783–6794, Online. Association for Computational Linguistics.

Bowei Zou, Qiaoming Zhu, and Guodong Zhou. 2015. Negation and speculation identification in Chinese language. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1), pages 656–665.