Noname manuscript No.
(will be inserted by the editor)

SCTB-V2: The 2nd Version of the Chinese Treebank in the Scientific Domain

**Chenhui Chu*, Zhuoyuan Mao,
Toshiaki Nakazawa, Daisuke Kawahara, Sadao
Kurohashi**

Received: date / Accepted: date

Abstract Word segmentation, part-of-speech (POS) tagging, and syntactic parsing are three fundamental Chinese analysis tasks for Chinese language processing, which are also crucial for various downstream tasks such as machine translation and information extraction. To achieve high accuracy for these tasks, treebanks that contain sentences manually annotated with word segmentation, part-of-speech tags, and phrase structures are essential. Although there are large-scale Chinese treebanks in the news domain, such treebanks are unavailable in the scientific domain. This significantly limits the performance of Chinese language processing for scientific text. To address this problem, we annotate the 2nd version of the Chinese treebank in the scientific domain (SCTB-V2). SCTB-V2 contains $12,175$ sentences annotated with word segmentation, part-of-speech tags, and phrase structures. We conducted Chinese analyses and machine translation experiments on SCTB-V2. The results show the effectiveness of SCTB-V2. We release this treebank to promote scientific Chinese language processing research.[1]

Keywords Treebank · Chinese · Scientific Domain

1 Introduction

Treebanks are text corpora containing sentences manually annotated with part-of-speech (POS) and syntactic information. For languages that do not have word boundaries, such as Chinese and Japanese, word segmentation also should be annotated in treebanks. The Penn treebank (PTB) (Marcus et al, 1993) is the first well-known treebank in English. Since the release of PTB, treebanks have played an important role in promoting natural language processing (NLP) research. Inspired by PTB, researchers are also constructing treebanks in other languages. A representative example is the

* Corresponding author: Chenhui Chu
E-mail: chu@i.kyoto-u.ac.jp, Kyoto University

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?A%20Chinese%20Treebank%20in%20Scientific%20Domain%20%28SCTB%29

universal treebank (Nivre et al, 2016). For the Chinese language, there also exist some treebanks. The two most commonly used ones are the Penn Chinese treebank (CTB) (Xue et al, 2005) and the Peking University (PKU) treebank (Yu et al, 2003). The development of Chinese treebanks has significantly promote Chinese language processing. For instance, Chinese analyses of word segmentation, POS tagging, and syntactic parsing, the F-Measures on the 5th version of CTB (CTB5)[2] is around 98%, 94% (Shen et al, 2014), and 80% (Petrov and Klein, 2007), respectively.

Domain difference is one difficult problem in NLP. Because most existing treebanks such as PTB, CTB, and PKU are annotated with news text when using the models trained on these treebanks to analyze sentences in other distant domains, it is difficult to obtain satisfying results. In China, scientific documents have been produced with a remarkable speed. For instance, the worldwide share of patent documents and scientific papers from China was 30% in 2009 (1st rank in the world),[3] and 13% on average from 2011 to 2013 (2nd in the world) (Saka and Igami, 2015), respectively. This leads to the rapid need increase for Chinese scientific text analysis, including text mining, knowledge discovery, and machine translation (MT). Unfortunately, the Chinese analysis performance significantly decreases when using news domain Chinese analysis models to analyze scientific text. Based on our preliminary experiments (see details in Section 3.1), the F-Measures decrease to 90%, 78%, and 67% for word segmentation, POS tagging, and syntactic parsing, respectively. The low analysis accuracy could be an error propagated to downstream tasks such as data mining and MT.

To promote NLP research for Chinese in the scientific domain, we constructed the 1st version of the Chinese treebank in the scientific domain (SCTB-V1) previously (Chu et al, 2016). However, SCTB-V1 only contains $5,133$ sentences ($138,781$ words) (Chu et al, 2016), which is significantly smaller compared to CTB-5, which has $18k$ sentences. In this paper, we release the 2nd version of SCTB (SCTB-V2), which contains $12,175$ sentences ($328,562$ words). Note that SCTB-V2 is a superset of SCTB-V2 but is more than twice larger than SCTB-V1, and the original annotations of SCTB-V1 have been revised according to some new standards in SCTB-V2. We present the details of the treebank annotation process of SCTB-V2. We select raw sentences from Chinese scientific papers for both SCTB-V1 and SCTB-V2. Our annotation standards essentially follow that of CTB (Xue et al, 2005). However, for Chinese word segmentation, we adopt the character-level POS pattern-based standard (Shen et al, 2016) to address the inconsistency and data sparsity problems of CTB. In addition, we design specific rules for expressions that are not covered by the CTB standard (Xue et al, 2005) but are frequently used in scientific documents, such as terminologies, formulas, and citations.

In order to investigate the usefulness of SCTB-V2, we first conducted instinct Chinese analysis experiments, including word segmentation, POS tagging, and syntactic parsing. The results verify that SCTB-V2 can significantly boost Chinese analysis by 2.09%, 4.76%, and 6.84% absolute F-Measure improvements compared to SCTB-V1, for word segmentation, POS tagging, and syntactic parsing, respectively.

[2] https://catalog.ldc.upenn.edu/LDC2005T01

[3] Statistics from Japan Patent Office.

In addition, we conducted extrinsic MT experiments on both Chinese-to-Japanese and Chinese-to-English on the scientific paper domain ASPCE-CJ and patent domain NTCIR-CE MT tasks. MT results also show that the SCTB-V2 significantly outperforms SCTB-V1.

The contributions of this paper are two-fold:

- We newly annotate SCTB-V2 with 12,175 sentences, whereas our previous SCTB-V1 only contains 5,133 sentences.
- Both instinct Chinese analysis and extrinsic MT experiments verify the effectiveness of SCTB-V2 compared to our previous work on SCTB-V1.

2 Treebank Annotation

Here, we present detailed information about our annotation, including raw sentence selection, annotation standards, and the real process.

2.1 Raw Sentence Selection

Japan Science and Technology Agency (JST) provides us with the National Science Library, Chinese Academy of Sciences (LCAS) corpus. Chinese scientific papers of many different sub-domains such as life science, computer science, biology, and chemistry are collected in the LCAS corpus. In addition, JST manually translated 780k Chinese abstracts of Chinese scientific papers in the LCAS corpus into Japanese. English translations are also available for a large number of them. For the LCAS corpus, we randomly selected the raw Chinese sentences that have both Japanese and English translations, leaving the possibility to extend our treebank to a trilingual one further.

2.2 Annotation Standards

Words are defined according to morphology analysis in previous segmentation standards, such as (Huang et al, 1996; Xia et al, 2000; Duan et al, 2003). However, this can cause data sparseness and inconsistency problems. For instance, according to the previous standards for segmentation, as "业 (industry)" has the characteristics of a bound morpheme and cannot be a word by itself, both "医药 (medicine)" and "医药业 (medicine industry)" in Figure 1 can be single words. This not only causes the segmentation inconsistency of "医药 (medicine)" but also leads to the sparsity of both words. Therefore, the character-level POS pattern-based Chinese word segmentation standard (Shen et al, 2016) is adopted in this paper. This standard can capture Chinese characters' grammatical roles inside words. In this standard, if a meaningful disyllabic string meets a predefined character-level POS pattern, we treat it as a word. For instance, as "医药 (medicine)" belongs to the "noun + noun" pattern it will be treated as one word, and thus "医药业 (medicine industry)" should be segmented into "医药 (medicine)" and "业 (industry)."

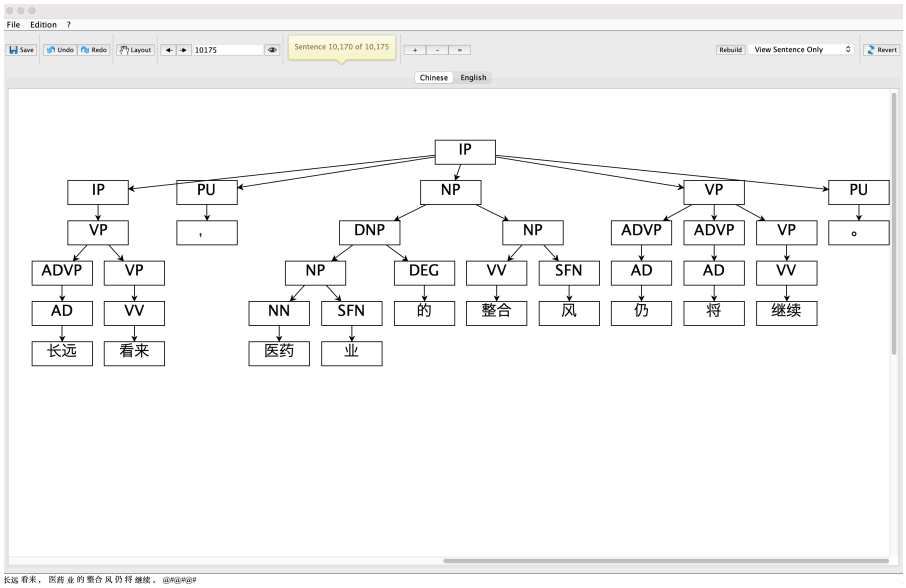

Fig. 1 An annotation interface screenshot for a Chinese sentence "长远 (long term) /看来 (see from) /，/医药 (medicine) /业 (industry) /的 ('s) /整合 (integrated) /风 (wind) /仍 (still) /将 (will) /继续 (continue)/。" (words are shown in the bottom boxes, POS tags are shown in the pre-terminal boxes, and phrasal constituents are shown in the upper boxes).

Essentially, we follow the CTB POS standard (Xue et al, 2005). In order to tag the bound morphemes, we further use six additional tags following (Shen et al, 2016). Among the six tags, three are for suffixes, namely, "SFA" (adjectival suffix), "SFN" (nominal suffix), and "SFV" (verbal suffix); and three are for prefixes, namely, "PFA" (adjectival prefix), "PFN" (nominal prefix), and "PFV" (verbal prefix). For instance, the tag of "业 (industry)" is "SFN."

For phrase structure annotation, we follow the standard of CTB (Xue et al, 2005). However, for single words in previous segmentation standards that are annotated as two words in our standard, they are combined into a single phrase structure constituent. For instance, in Figure 1 we combine "医药_NN (medicine) /业_SFN (industry)" into an NP (noun phrase).

Due to the scientific domain, there are various specific expressions that cannot be covered by the CTB standard (Xue et al, 2005), such as terminologies, formulas, and citations. We design specific rules for those expressions. In addition, we updated some specific rules from SCTB-V1 to cover the new linguistic phenomena in SCTB-V2. The following are the major POS tagging standard updates from SCTB-V1 to SCTB-V2.

- Keep disyllabic era names as single tokens (e.g., "唐代 (Tang dynasty)").
- When there is a proper noun in a scientific name of creatures, segment it (e.g., "埃及 (Aedes) /伊蚊 (aegypti)").

- Treat superscript numbers as single tokens (e.g., "10^6").
- Do not segment the string if it is a combination of foreign characters and numbers/ symbols/transliterations (e.g., "卡拉OK (Karaoke)").
- Treat abbreviation nouns including numbers as single tokens (e.g., "乙肝 (Hepatitis B)").
- When there is a proper noun in a telescopic compound, segment it (e.g., "中 (Chinese) /西 (Western) /医 (medicine)").
- Keep nounized disyllabic patterns of "CC+NN" and "CC+NNB" as as single tokens (e.g., 一员 (a member)")).
- Treat unit symbols consisting of the alphabets as single tokens (e.g., "１３.５ / ｍｇ").

The following are the major POS tagging and parsing standard updates from SCTB-V1 to SCTB-V2.

- Let POS tags of temporal nouns be NT (e.g., "目前_NT (currently)").
- Let POS tags of localizers be either LC or NN (e.g., "这_DT (this) /个_M (piece) /问题_NN (problem) /上_LC (at)").
- Label temporal noun phrases consisting of cardinals and quantifiers as NP (e.g., "NP (21_CD /世纪_M (century))").

More detailed rules and our segmentation standard will be released in the future.

2.3 Annotation Process

In order to annotate SCTB, the SynTree toolkit[4] was used. SynTree provides an annotator-friendly graphical interface to annotate phrase structures. To annotate word segmentation, POS tags, and phrase structures, annotators can simply drag and edit boxes containing either words, POS tags, or phrasal constituents. An example of the SynTree toolkit is shown in Figure 1. To make the annotators conduct annotations more easily, we got feedback from them and further improved the toolkit according to the feedback throughout the entire annotation process.

Two annotators conducted the annotation: FH and TU. FH was a one-year experienced annotator, but TU had no experience. Therefore, we asked FH not only to train TU but also to review FH's annotation. To speed up the annotation, we segmented, pos-tagged, and parsed the Chinese sentences with a baseline Chinese analysis system (see Section 3.1). Based on the results from the baseline system, FH and TU manually revised the errors using the SynTree toolkit.[5] Different sentences were annotated by FH and TU. After the annotation, FH further reviewed and revised the sentences annotated TU. Different from a conventional way, inter-annotator agreements were calculated based on the review and revision results. We compared the sentences without and with the review and revision to get the inter-annotator agreements. The annotation agreements were 98.95%, 97.78%, and 95.05% for word segmentation, POS tagging, and syntactic parsing, respectively.

[4] http://syntree.github.io/index.html

[5] As both of the two workers were well trained through the entire two years annotation period, we did not observe biases introduced by the baseline systems in the final SCTB-V2.

In the release of SCTB-V2, we have annotated and reviewed $12,175$ sentences ($328,562$ words). With the above efforts, it still took two years to finish the annotation. The average speed of annotation was around 5 Chinese sentences per hour and per person.

3 Experiments

In order to show the effectiveness of SCTB-V2, both Chinese analysis and MT experiments were conducted. For Chinese analysis, we conducted word segmentation, POS tagging, and syntactic parsing experiments. For MT, which is an important downstream task for Chinese analysis, we conducted experiments on both the scientific paper and patent domains for both Chinese-to-Japanese and Chinese-to-English translations.

3.1 Chinese Analysis Experiments

For word segmentation and POS tagging experiments, we used KyotoMorph[6] (Shen et al, 2014). For syntactic parsing experiments, we used Berkeley parser[7] proposed by Petrov and Klein (2007) and Berkeley neural parser[8] proposed by Kitaev et al (2019).

For Chinese analysis experiments, SCTB-V2 was split into training, validation, and testing sets with $10,175$, $1,000$, and $1,000$ sentences, respectively. The following three settings were compared:

- Baseline: We trained the Chinese analyzers on the union of two baseline treebanks.[9] The first one is CTB5, which has 18k news domain sentences. Instead of using the original CTB5, we used the version re-annotated according to character-level POS patterns (Shen et al, 2016), which follows the same standard as SCTB. The second one is an in-house treebank, which contains 10k sentences mostly in the NLP domain. This treebank also follows our word segmentation standard.
- Baseline+SCTB-V1: Used the Baseline treebanks together with the training split in SCTB-V1 (Chu et al, 2016) for training the Chinese analyzers.
- SCTB-V2: Used the training split in SCTB-V2 for training the Chinese analyzers.
- Baseline+SCTB-V2: Used the Baseline treebanks together with the training split in SCTB-V2 for training the Chinese analyzers.

The analysis results were reported on the testing split in SCTB-V2. We conducted significance tests with the bootstrap re-sampling method proposed by Koehn (2004).

The results of word segmentation, joint segmentation and POS tagging are shown in Tables 1 and 2, respectively. We can see that a large margin of improvement is

[6] https://bitbucket.org/msmoshen/kyotomorph-beta

[7] https://github.com/slavpetrov/berkeleyparser

[8] https://github.com/nikitakit/self-attentive-parser

[9] Preliminary experiments show that the union is better than using one of them only.

System	Precision	Recall	F-Measure
Baseline	90.21	90.72	90.46
Baseline+SCTB-V1	95.04	95.65	95.35
SCTB-V2	**97.28**	**97.59**	**97.44**†‡
Baseline+SCTB-V2	96.80	97.20	97.00†‡

Table 1 Word segmentation results ("†" and "‡" indicate that the result is significantly better than "Baseline" and "Baseline+SCTB-V1" at $p < 0.01$, respectively).

System	Precision	Recall	F-Measure
Baseline	78.61	79.06	78.83
Baseline+SCTB-V1	87.89	88.45	88.17
SCTB-V2	**92.78**	**93.08**	**92.93**†‡
Baseline+SCTB-V2	91.78	92.17	91.97†‡

Table 2 Results for joint segmentation and POS tagging ("†" and "‡" indicate that the result is significantly better than "Baseline" and "Baseline+SCTB-V1" at $p < 0.01$, respectively).

observed by comparing Baseline+SCTB-V1 to Baseline, i.e., 4.89% and 9.34% F-Measure for word segmentation, and joint segmentation and POS tagging, respectively. SCTB-V2 further significantly boosts the performance from Baseline+SCTB-V1, with 2.09% and 4.76% F-Measure improvements in word segmentation, joint segmentation and POS tagging, respectively. However, Baseline+SCTB-V2 slightly decreases the performance compared to SCTB-V2. We think the reason is due to the difference between the annotation standards of Baseline and SCTB-V2, i.e., adding the Baseline treebank belonging to a different standard cannot improve the performance tested on SCTB-V2.

Syntactic parsing results of Berkeley parser and Berkeley neural parser are shown in Table 3 and 4, respectively. The Evalb toolkit[10] was used to calculate parsing accuracy. Note that originally, Evalb was designed for English parsing accuracy calculation. Therefore, without the same word segmentation as the ground-truth data, Evalb cannot calculate parsing accuracy. Due to this, we report results with ground-truth word segmentation in both Tables 3 and 4. From the results of the Berkeley parser reported in Table 3, we can see that the improvement from Baseline to Baseline+SCTB-V1 is significant, where the F-Measure gap is 9.17%; SCTB-V2 further significantly boots the performance from Baseline+SCTB-V1, with a 6.84% F-Measure improvement; Similarly, using Baseline+SCTB-V2 leads to a slight decrease in parsing performance compared to SCTB-V2. We think the reason is the same as the one for the performance decrease of word segmentation, joint segmentation and POS tagging, which is due to the difference between the annotation standards of Baseline and SCTB-V2. From the results of Berkeley neural parser reported in Table 4, we can see that the same trends as in Table 3 that Baseline+SCTB-V1 outperforms Baseline with a large margin of 7.77% F-Measure; SCTB-V2 further significantly improves the performance from Baseline+SCTB-V1, with a 5.08% F-Measure improvement; Baseline+SCTB-V2 shows a slight decrease in parsing performance compared to

[10] http://nlp.cs.nyu.edu/evalb/

SCTB-V2. Comparing the results of the Berkeley parser with those of the Berkeley neural parser, we also see significant improvements by Berkeley neural parser, showing the same observations for the effectiveness of using pre-trained language models in syntactic parsing as reported in (Kitaev et al, 2019).

System	Precision	Recall	F-Measure
Baseline	72.26	63.57	67.64
Baseline+SCTB-V1	80.61	73.36	76.81
SCTB-V2	**83.91**	**83.38**	**83.65**†‡
Baseline+SCTB-V2	83.76	81.26	82.49†‡

Table 3 Syntactic parsing results based on ground-truth segmentation of Berkeley parser. ("†" and "‡" indicate that the result is significantly better than "Baseline" and "Baseline+SCTB-V1" at $p < 0.01$, respectively).

System	Precision	Recall	F-Measure
Baseline	83.96	69.24	75.89
Baseline+SCTB-V1	87.79	79.90	83.66
SCTB-V2	**90.34**	**87.20**	**88.74**†‡
Baseline+SCTB-V2	90.20	87.15	88.65†‡

Table 4 Syntactic parsing results based on ground-truth segmentation of Berkeley neural parser ("†" and "‡" indicate that the result is significantly better than "Baseline" and "Baseline+SCTB-V1" at $p < 0.01$, respectively).

To understand the improvements deeply, we further the manually analyzed results. According to our analyses, most results are improved because of the scientific domain knowledge in SCTB-V2. An improved example is shown in Figure 2. Note that the syntactic parsing results of this example were obtained by the Berkeley parser. Baseline segments "艾滋病患者 (AIDS patient)" into "艾滋 (AIDS) /病患 (patient) /者 (people)" incorrectly, because it lacks the knowledge that "艾滋病 (AIDS disease)" is a medical term that should be segmented separately instead of combing to further coming words. The segmentation error is also propagated to POS tagging and syntactic parsing. In contrast, with SCTB-V2 correctly segments it into "艾滋 (AIDS) /病 (disease) /患者 (patient)," which also improves the POS tagging and parsing accuracy.

To learn the relationship between the number of sentences being annotated and the performance of Chinese analysis in detail, we further conducted experiments training Chinese analyzers with an incremental number of sentences from SCTB-V2. In detail, we increasingly used $2,000$ sentences from SCTB-V2 to train the analyzers. Results are shown in Figure 3. Note that the syntactic parsing results were obtained by the Berkeley parser. We find that the improvement by adding more sentences is most significant for parsing, followed by POS tagging and segmentation; for all word segmentation, POS tagging, and syntactic parsing, the improvements in accuracy slow down using more sentences to train the analyzers.

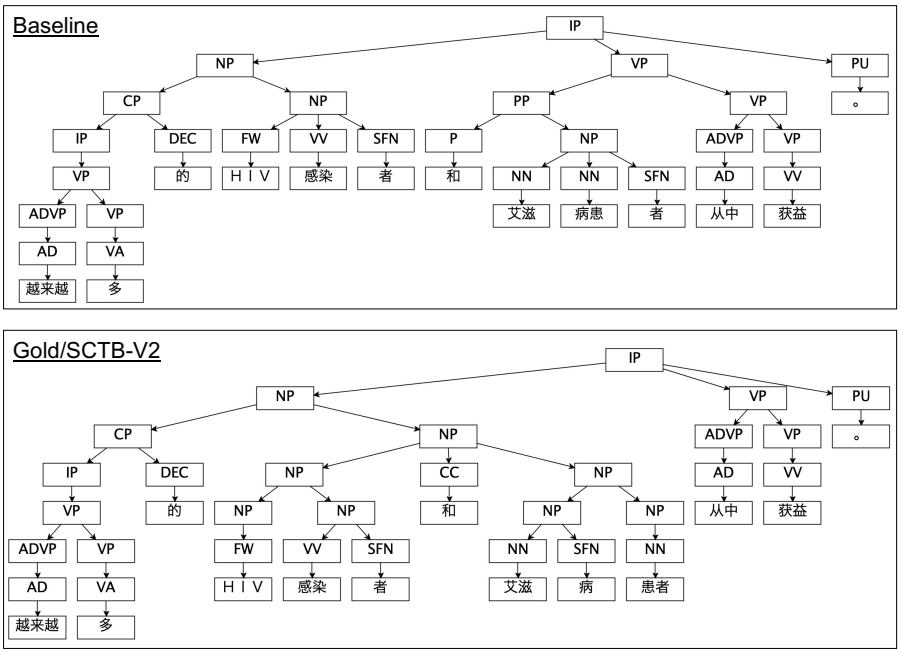

Fig. 2 A example comparison between baseline and SCTB-2 of Chinese analysis results of a Chinese sentence "越来越 (more) /多 (more) /的 (of) /ＨＩＶ (HIV) /感染 (infection) /者 (people) /和 (and) /艾滋 (AIDS) /病 (disease) /患者 (patient) /从中 (from) /获益 (benefit) /。"

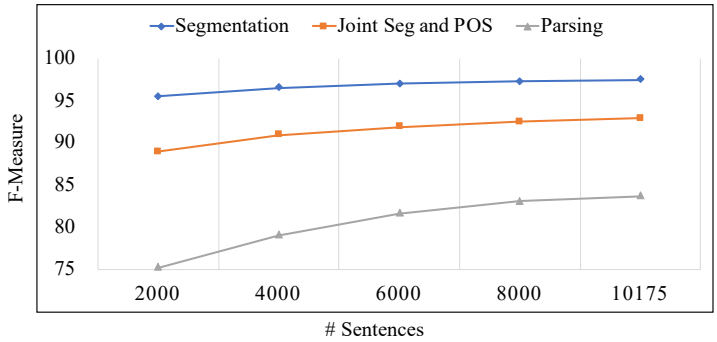

Fig. 3 Results for Chinese analysis using incremental numbers of sentences selected from SCTB-V2 for training the Chinese analyzers.

3.2 MT Experiments

Chinese-to-Japanese MT experiments were conducted on ASPEC-CJ corpus[11] (Nakazawa et al, 2016), which is a scientific domain corpus. The ASPEC-CJ corpus has been

[11] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

System	ASPEC-CJ	NTCIR-CE
Baseline	34.81	31.65
Baseline+SCTB-V1	35.06	31.18
SCTB-V2	35.28	31.39
Baseline+SCTB-V2	**35.38**†‡	**32.25**†‡

Table 5 MT results (BLEU-4 scores) on ASPEC-CJ and NTCIR-CE ("†" and "‡" indicate that the result is significantly better than "Baseline" and "Baseline+SCTB-V1" at $p < 0.05$, respectively).

used for a shared task at the Workshop on Asian Translation (WAT)[12] (Nakazawa et al, 2021). This task contains $672,315$ training, $2,090$ validation, and $2,107$ testing sentences, respectively. Chinese-to-English MT experiments were conducted on the NTCIR-CE corpus, which belongs to the patent domain. This corpus has been used for a shared task at the NTCIR-10 workshop[13] (Goto et al, 2013). This task contains $1,000,000$ training, $2,000$ validation, and $2,000$ testing sentences, respectively.

We conducted tree-to-string experiments, where Chinese is parsed to syntactic trees while Japanese/English sentences are used as they are. KyotoMorph was used to word segment Chinese sentences, and Berkeley parser was used for joint POS tagging and syntactic parsing. The syntactic parsing trees were further binarized to extract translation rules better. Same to the Chinese analysis experiments, we compared the settings of "Baseline," "Baseline+SCTB-V1," "SCTB-V2," and "Baseline+SCTB-V2" in Section 3.1 for MT. To segment Japanese sentences, JUMAN[14] (Kurohashi et al, 1994) was used. We also tokenized English sentences with a tokenization script in Moses (Koehn et al, 2007).

For MT experiments, we used the tree-to-string neural MT toolkit provided by Chen et al (2017).[15] All the hyper-parameters were tuned on the ASPEC-CJ task and directly reused on the NTCIR-CE task. Specifically, the maximum length for each sentence was set to 75 words. The word embedding dimension and the hidden state dimension were set to 512 and 768 for the tree-to-string neural model, respectively. We used Adadelta (Zeiler, 2012) for optimization with a batch size of 16 and a learning rate of 0.0005. Each experiment was trained on a single TITAN Xp GPU card, and the model was validated every $1,000$ training step. The training was early stopped if no improvement of the validation loss was observed within 50 checkpoints. We ran each setting 3 times and reported the average BLEU score (Papineni et al, 2002) for evaluation.

MT results are shown in Table 5. We also conducted significance tests for MT using the bootstrap re-sampling method proposed by Koehn (2004). We find that there are also significant improvements in MT performance with SCTB-V2. Besides the language pair and domain difference, similar improvements can be seen in the results for both ASPEC-CJ and NTCIR-CE. Especially, Baseline+SCTB-V2 significantly outperforms both Baseline and Baseline+SCTB-V1. Different from the Chinese analysis results, Baseline+SCTB-V2 performs better than SCTB-V2 in MT. We suspect

[12] http://orchid.kuee.kyoto-u.ac.jp/WAT/

[13] http://ntcir.nii.ac.jp/PatentMT-2/

[14] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[15] https://github.com/howardchenhd/Syntax-awared-NMT

Source	气雾剂容器 2 设置有阀 1 1 ， 阀 1 1 与喷嘴一体形成。
Reference	Aerosol container 2 is provided with a valve 11 which is formed integrally with a spray nozzle.
Baseline	The valve 1 is provided with a valve 11 , and the valve 11 is formed integrally with the nozzle.
Baseline +SCTB-V2	The aerosol container 2 is provided with a valve 11 , which is integrally formed with the nozzle.

Table 6 An improved MT example.

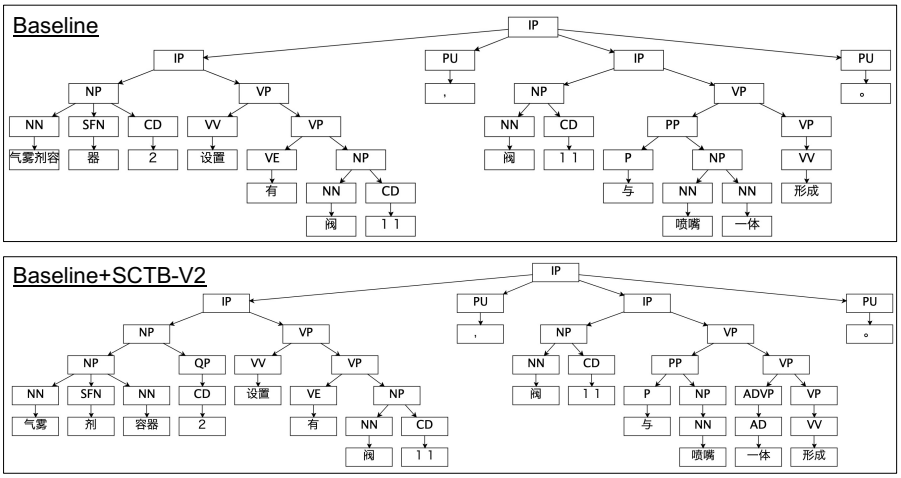

Fig. 4 The analysis results for the Chinese source sentence in Table 6 “气雾 (aerosol) /剂 (drug) /容器 (container) / 2 /设置 (provide) /有 (with) /阀 (valve) / 1 1 /， /阀 / 1 1 /与 (is) /喷嘴 (nozzle) /一体 (integrally) /形成 (formed) /。”

the reason for this is that combining the Baseline and SCTB-V2 treebanks makes Chinese analysis more robust for MT.

We also investigated the results to understand where translation improvements come from. Based on our investigation, we noticed that most improvements come from the Chinese analysis improvements. An NTCIR-CE MT example improved by SCTB-V2 is shown in Table 6. We can see that “气雾剂容器 (aerosol container)” is incorrectly translated into “the valve” by Baseline. This happens because the Baseline analyzes “气雾剂容器 (aerosol container)” as “气雾剂容/器,” where the segmentation is semantically incorrect, as shown in Figure 4. There is another analysis error for “一体 (integrally) /形成 (formed)” by Baseline, where “一体 (integrally)” is analyzed as a noun “one” and the parsing result is also wrong correspondingly. Fortunately, the neural MT model is not affected by this wrong analysis result.

4 Related Work

CTB has been continuously annotated, and the latest version is CTB9.[16] CTB9 has $132,076$ sentences, which is significantly larger than CTB5. Two other treebanks are available for Chinese besides CTB (Xue et al, 2005). The first one is the PKU Chinese treebank, which takes a two-step annotation process 1) word segmentation and POS tags (Yu et al, 2003), and 2) syntactic parsing (Qiu et al, 2014). The second one is the Harbin Institute of Technology (HIT) treebank, whose syntactic parsing annotation is based on dependency structures (Che et al, 2012). Note that CTB, PKU, and HIT treebanks adopt different annotation standards. All the sentences in these three treebanks belong to the news domain. Raw sentences of the CTB treebank were selected from various news agencies, including Xinhua newswire, Hong Kong newswire, and People's Daily. Raw sentences of the PKU and HIT treebank were selected from the People's Daily newswire. Recently, a Chinese treebank in the literature domain has also been constructed (Hu et al, 2020). Therefore, SCTB is the only publicly released scientific domain Chinese treebank.

Two types of syntactic grammar are used in treebanks: phrase and dependency structures. In SCTB, phrase structures were adopted similar to CTB (Xue et al, 2005). The reason for this is that we can easily convert phrase structures to dependency structures according to predefined head rules available in the Penn2Malt toolkit.[17] Qiu et al (2014) also proposed a multi-view including both phrase and dependency structures for treebanking.

With a significant need increase in multilingual NLP, multilingual treebanks have been developed recently. The universal dependency treebank[18] (Nivre et al, 2016) and the Asian language treebank (Thu et al, 2016) is two representatives. With multilingual treebanks, multilingual syntactic parsing becomes possible and shared tasks have been organized for this (Zeman et al, 2018). We selected the raw sentences of SCTB-V2 from a parallel corpus. Therefore, we can further develop SCTB-V2 to a trilingual treebank by annotating the corresponding Japanese and English parallel sentences.

5 Conclusion

This paper presented detailed information on the annotation of SCTB-V2: the 2nd version of the Chinese treebank in the scientific domain. Both Chinese analysis and MT experiments showed that SCTB-V2 performs significantly better than both a baseline treebank and our previous SCTB-V1. In future work, we plan to annotate Japanese further and English translations of the sentences in SCTB-V2 to make SCTB be trilingual.

[16] https://catalog.ldc.upenn.edu/LDC2016T13

[17] http://stp.lingfil.uu.se/∼nivre/research/Penn2Malt.html

[18] http://universaldependencies.org

Acknowledgements

This work was supported by "Project on Practical Implementation of Japanese to Chinese-Chinese to Japanese Machine Translation,"[19] JST. We sincerely thank Ms. Fumio Hirao and Mr. Teruyasu Ueki, who annotated SCTB-V2. We are appreciated Mr. Frederic Bergeron for his development of the SynTree toolkit to speed up the annotation process. Finally, we want to thank Dr. Mo Shen for valuable discussions regarding annotation standards.

References

Che W, Li Z, Liu T (2012) Chinese dependency treebank 1.0. In: Linguistic Data Consortium

Chen H, Huang S, Chiang D, Chen J (2017) Improved neural machine translation with a syntax-aware encoder and decoder. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp 1936–1945, URL http://aclweb.org/anthology/P17-1177

Chu C, Nakazawa T, Kawahara D, Kurohashi S (2016) SCTB: A Chinese treebank in scientific domain. In: Proceedings of the 12th Workshop on Asian Language Resources (ALR12), The COLING 2016 Organizing Committee, Osaka, Japan, pp 59–67, URL https://aclanthology.org/W16-5407

Duan H, Bai X, Chang B, Yu S (2003) Chinese word segmentation at peking university. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Association for Computational Linguistics, Sapporo, Japan, pp 152–155, DOI 10.3115/1119250.1119272, URL http://www.aclweb.org/anthology/W03-1722

Goto I, Chow KP, Lu B, Sumita E, Tsou BK (2013) Overview of the patent machine translation task at the ntcir-10 workshop. In: Proceedings of the 10th NTCIR Conference, National Institute of Informatics (NII), Tokyo, Japan, pp 260–286, URL http://dblp.uni-trier.de/db/conf/ntcir/ntcir2013.html#GotoCLST13

Hu H, Li Y, Patterson Y, Tian Z, Zhang Y, Zhou H, Kuebler S, Lin CJC (2020) Building a treebank for Chinese literature for translation studies. In: Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories, Association for Computational Linguistics, Düsseldorf, Germany, pp 18–30, DOI 10.18653/v1/2020.tlt-1.2, URL https://aclanthology.org/2020.tlt-1.2

Huang CR, Chen KJ, Chang LL (1996) Segmentation standard for chinese natural language processing. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 2, Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '96, pp 1045–1048, DOI 10.3115/993268.993362, URL http://dx.doi.org/10.3115/993268.993362

Kitaev N, Cao S, Klein D (2019) Multilingual constituency parsing with self-attention and pre-training. In: Proceedings of the 57th Annual Meeting of the As-

[19] https://jipsti.jst.go.jp/jazh_zhja_mt/en.html

sociation for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 3499–3505, DOI 10.18653/v1/P19-1340, URL https://aclanthology.org/P19-1340

Koehn P (2004) Statistical significance tests for machine translation evaluation. In: Lin D, Wu D (eds) Proceedings of EMNLP 2004, Association for Computational Linguistics, Barcelona, Spain, pp 388–395

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics, Prague, Czech Republic, pp 177–180, URL http://www.aclweb.org/anthology/P/P07/P07-2045

Kurohashi S, Nakamura T, Matsumoto Y, Nagao M (1994) Improvements of Japanese morphological analyzer JUMAN. In: Proceedings of the International Workshop on Sharable Natural Language, pp 22–28

Marcus MP, Marcinkiewicz MA, Santorini B (1993) Building a large annotated corpus of english: The penn treebank. Computational Linguistics 19(2):313–330, URL http://dl.acm.org/citation.cfm?id=972470.972475

Nakazawa T, Yaguchi M, Uchimoto K, Utiyama M, Sumita E, Kurohashi S, Isahara H (2016) Aspec: Asian scientific paper excerpt corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

Nakazawa T, Nakayama H, Ding C, Dabre R, Higashiyama S, Mino H, Goto I, Pa Pa W, Kunchukuttan A, Parida S, Bojar O, Chu C, Eriguchi A, Abe K, Oda Y, Kurohashi S (2021) Overview of the 8th workshop on Asian translation. In: Proceedings of the 8th Workshop on Asian Translation (WAT2021), Association for Computational Linguistics, Online, pp 1–45, DOI 10.18653/v1/2021.wat-1.1, URL https://aclanthology.org/2021.wat-1.1

Nivre J, de Marneffe MC, Ginter F, Goldberg Y, Hajic J, Manning CD, McDonald R, Petrov S, Pyysalo S, Silveira N, Tsarfaty R, Zeman D (2016) Universal dependencies v1: A multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France, pp 1659–1666

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318, DOI 10.3115/1073083.1073135, URL https://aclanthology.org/P02-1040

Petrov S, Klein D (2007) Improved inference for unlexicalized parsing. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Association for Computational Linguistics, Rochester, New York, pp 404–411, URL http://www.aclweb.org/anthology/N/N07/N07-1051

Qiu L, Zhang Y, Jin P, Wang H (2014) Multi-view chinese treebanking. In: Proceedings of COLING 2014, the 25th International Conference on Computational Lin-

guistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp 257–268, URL http://www.aclweb.org/anthology/C14-1026

Saka A, Igami M (2015) Benchmarking scientific research 2015. Ministry of Education, Culture, Sports, Science and Technology, Japan, pp 1–172

Shen M, Liu H, Kawahara D, Kurohashi S (2014) Chinese morphological analysis with character-level pos tagging. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Baltimore, Maryland, pp 253–258, URL http://www.aclweb.org/anthology/P14-2042

Shen M, Wingmui L, Choe H, Chu C, Kawahara D, Kurohashi S (2016) Consistent word segmentation, part-of-speech tagging and dependency labelling annotation for chinese language. In: Proceedings of the 26th International Conference on Computational Linguistics, Association for Computational Linguistics, Osaka, Japan

Thu YK, Pa WP, Utiyama M, Finch A, Sumita E (2016) Introducing the asian language treebank (alt). In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France

Xia F, Palmer M, Xue N, Okurowski ME, Kovarik J, dong Chiou F, Huang S, Kroch T, Marcus M (2000) Developing guidelines and ensuring consistency for chinese text annotation. In: In Proceedings of the Second Language Resources and Evaluation Conference

Xue N, Xia F, Chiou Fd, Palmer M (2005) The penn chinese treebank: Phrase structure annotation of a large corpus. Natural Language Engineering 11(2):207–238, DOI 10.1017/S135132490400364X, URL http://dx.doi.org/10.1017/S135132490400364X

Yu S, Duan H, Swen B, Chang B (2003) Specification for corpus processing at peking university: Word segmentation, POS tagging and phonetic notation. Journal of Chinese Language and Computing 13(2):121–158

Zeiler MD (2012) ADADELTA: an adaptive learning rate method. CoRR abs/1212.5701, URL http://arxiv.org/abs/1212.5701, 1212.5701

Zeman D, Hajič J, Popel M, Potthast M, Straka M, Ginter F, Nivre J, Petrov S (2018) CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, pp 1–21, DOI 10.18653/v1/K18-2001, URL https://aclanthology.org/K18-2001