SCTB: A Chinese Treebank in Scientific Domain

Chenhui Chu,[1] **Toshiaki Nakazawa,**[1] **Daisuke Kawahara**[2] and **Sadao Kurohashi**[2]
[1]Japan Science and Technology Agency
[2]Graduate School of Informatics, Kyoto University
{chu,nakazawa}@pa.jst.jp {dk,kuro}@i.kyoto-u.ac.jp

Abstract

Treebanks are curial for natural language processing (NLP). In this paper, we present our work for annotating a Chinese treebank in scientific domain (SCTB), to address the problem of the lack of Chinese treebanks in this domain. Chinese analysis and machine translation experiments conducted using this treebank indicate that the annotated treebank can significantly improve the performance on both tasks. This treebank is released to promote Chinese NLP research in scientific domain.

1 Introduction

A treebank is a text corpus consisting of usually thousands to tens of thousands of sentences annotated with linguistic knowledge such as segmentation, part-of-speech (POS) tags and syntactic structures. From the initial release of the first treebank of the Penn treebank (PTB) (Marcus et al., 1993), treebanking has remarkably promoted the research of statistical natural language processing (NLP). Inspired by the success of the English treebank, treebanks for other languages have also been constructed or under construction (Nivre et al., 2016). For Chinese, there are several existing treebanks such as the widely used Penn Chinese treebank (CTB) (Xue et al., 2005), and the Peking University (PKU) treebank (Yu et al., 2003). Chinese language processing has been significantly developed with these treebanks. For example, the F-Measures of Chinese analysis on the benchmark data set CTB version 5 (CTB5)[1] has achieved about 98% for segmentation, 94% for POS tagging (Shen et al., 2014), and 80% for syntactic parsing (Petrov and Klein, 2007).

One difficulty of statistical NLP is the domain diversity. As most treebanks such as the PTB, CTB and PKU are constructed mainly in news domain, the performance is not satisfied when analyzing sentences in other distant domains using the models trained on these treebanks. In China, the number of scientific documents has been remarkably increased. For example, the worldwide share of patent documents has increased to 30% (worldwide rank 1) in 2009,[2] and the worldwide share of scientific papers has increased to 13% on the average of 2011-2013 (worldwide rank 2) (Saka and Igami, 2015). Therefore, the needs for scientific domain text analyzing such as text mining, knowledge discovery, and translating scientific documents to other languages are increasing. However, when applying the Chinese analysis models trained on different domains to scientific domain, the F-Measures of various analysis tasks dramatically decrease to 90% for segmentation, 78% for POS tagging, and 70% for syntactic parsing (Section 3.1). This level of low accuracy analysis could significantly affect the performance of downstream applications such as text mining and machine translation (MT).

Motivated by this, we decide to construct a Chinese treebank in the scientific domain (SCTB) to promote Chinese NLP research in this domain. This paper presents the details of our treebank annotation process and the experiments conducted on the annotated treebank. The raw sentences are selected from Chinese scientific papers. Our annotation process follows that of CTB (Xue et al., 2005) with an exception of the segmentation standard. We apply a Chinese word segmentation standard based on character-level POS patterns (Shen et al., 2016), aiming to circumvent inconsistency and address data

---

[1]https://catalog.ldc.upenn.edu/LDC2005T01
[2]Statistics from Japan Patent Office.

Figure 1: A screenshot of the annotation interface containing an annotation example of a Chinese sentence "烟草 (tobacco) /使用 (use) /是 (is) /当今 (nowadays) /世界 (world) /最大 (biggest) /的 ('s) /可 (can) /预防 (prevention) /死因 (cause of death) /，/烟草 (tobacco) /使用 (use) /者 (person) /中 (among) /近 (about) /一半 (half) /将 (will) /死于 (die) /烟草 (tobacco) /使用 (use) /。" (the bottom boxes contain words, the pre-terminal boxes contain POS tags, while the upper boxes contain phrasal constituents).

sparsity of the annotated treebank. As the first version of release, we finished the annotation of 5,133 sentences (138,781 words).[3] To verify the effectiveness of the annotated SCTB, we conducted both instinct Chinese analysis experiments of segmentation, POS tagging and syntactic parsing, and extrinsic MT experiments on Chinese-to-Japanese and Chinese-to-English directions. Experimental results show that the annotated SCTB can significantly improve both Chinese analysis and MT performance.

2 Treebank Annotation

We annotate segmentation, POS tags and phrase structures for sentences in scientific domain. In this section, we describe the details for the annotation.

2.1 Raw Sentence Selection

The raw Chinese sentences for the treebank annotation are selected from the LCAS (National Science Library, Chinese Academy of Sciences) corpus provided by Japan Science and Technology Agency (JST). The LCAS corpus consists of Chinese scientific papers of various scientific subdomains. From this corpus, 780k abstracts were manually translated from Chinese to Japanese by JST (most of them also contain English translations). We randomly selected the raw sentences from the parallel part of the LCAS corpus, aiming for not only improving Chinese analysis but also multilingual NLP.

2.2 Annotation Standard

Conventional segmentation standards (Huang et al., 1996; Xia et al., 2000; Duan et al., 2003) define words based on the analysis of morphology, which could lead to two problems: inconsistency and data sparsity. For example, based on the conventional segmentation standards, both "使用 (use)" and "使用者 (user/use person)" in Figure 1 are one words, because "者 (person)" is a bound morpheme that cannot form a word itself. This leads to the inconsistent segmentation of "使用 (use)", and also makes

---

both words sparse. In this work, we adopt the Chinese word segmentation standard based on character-level POS patterns (Shen et al., 2016), which captures the grammatical roles of Chinese characters inside words. In our standard, we only treat a meaningful disyllabic string as a word if it falls into one predefined character-level POS patterns. For example, "使用 (use)" is one word as it belongs to the "verb + verb" pattern, and thus "使用者 (user/use person)" should be segmented into "使用" and "者 (person)".

Our POS standard essentially follows the one used in CTB (Xue et al., 2005). In order to tag the bound morphemes in conventional segmentation standards, we add six new tags into the tag set following (Shen et al., 2016), three for suffixes: "SFN" (nominal suffix), "SFA" (adjectival suffix), and "SFV" (verbal suffix); and three for prefixes: "PFN" (nominal prefix), "PFA" (adjectival prefix), and "PFV" (verbal prefix). For example, "者 (person)" is tagged with "SFN".

Our phrase structure annotation standard also follows that of CTB (Xue et al., 2005). For the words that should be one word according to the conventional segmentation standards, we combine them into one constituent in the phrase structure level. For example, "使用_VV (use) /者_SFN (person)" is combined into an NP (noun phrase) in Figure 1.

As we are annotating scientific texts, there are many specific expressions such as terminologies, formulas, and citations, which have not been covered by the conventional standards (Xue et al., 2005). For these, we define specific rules in particular. We plan to release the details of these rules together with our segmentation guideline along with the treebank.

2.3 Annotation Process

We used the SynTree toolkit[4] as the annotation interface. SynTree is a graphical interface for phrase structure annotation. Users can perform all the segmentation, POS tag and phrase structure annotations in this interface via dragging and editing the boxes containing words, POS tags and phrasal constituents in a bottom-up manner. We also customized the toolkit based on the feedbacks of the annotators during the annotation process. Figure 1 is a screenshot of the annotation interface.

The annotation was performed by two annotators: H and U. Annotator H had one year annotation experience, while annotator U was fresh at the beginning of the annotation. Therefore, annotator H was also responsible for training annotator U and reviewing the annotation done by annotator U. To improve the efficiency of annotation, the raw sentences were firstly processed by a baseline system described in Section 3.1. The annotators did the annotation by revising the errors in the automatic analysis results made by the baseline system using the annotation interface. The two annotators were asked to annotate different sentences, respectively. After that annotator H was asked to review and revise the sentences annotated by annotator U. We calculated the inter-annotator agreements on the sentences before and after the review/revision, and found that the agreements for segmentation, POS tagging and parsing are 98.95%, 97.78%, and 95.05%, respectively.

We have finished the annotation and review for 5,133 sentences (138,781 words) at the end of August, 2016. It took us 6 months for this annotation, and the average annotation speed was about 5 sentences/hour per person.

3 Experiments

We conducted both instinct and extrinsic experiments to verify the effectiveness of the annotated treebank. The instinct experiments were the conventional Chinese analysis tasks including segmentation, POS tagging and syntactic parsing. For the extrinsic experiments, we selected MT as an application of the Chinese analysis tasks, and conducted MT experiments on both the Chinese-to-Japanese and Chinese-to-English directions in scientific paper and patent domains, respectively.

3.1 Analysis Experiments

We conducted segmentation, POS tagging and syntactic parsing experiments. Segmentation and POS tagging experiments were conducted using the Chinese analyzing tool KyotoMorph[5] proposed by Shen

---

[4]http://syntree.github.io/index.html
[5]https://bitbucket.org/msmoshen/kyotomorph-beta

System	Precision	Recall	F-Measure
Baseline	90.99	89.97	90.48
Baseline+SCTB	94.59	94.91	94.75†

Table 1: Word segmentation results ("†" indicates that the result is significantly better than "Baseline" at $p < 0.01$).

System	Precision	Recall	F-Measure
Baseline	78.21	77.33	77.77
Baseline+SCTB	84.88	85.17	85.03†

Table 2: Joint segmentation and POS tagging results ("†" indicates that the result is significantly better than "Baseline" at $p < 0.01$).

et al. (2014). Parsing was performed by the Berkeley parser[6] (Petrov and Klein, 2007). We compared the Chinese analysis performance of the Chinese analyzers trained on the following two settings.

- Baseline: Chinese analyzers trained on CTB5 containing 18k sentences in news domain, and a previously created in-house (mainly) NLP domain treebank of 10k sentences. Note that the Chinese word segmentation of the baseline treebanks originally follows the conventional segmentation standard (Xia et al., 2000), and we manually re-annotated them based on the character-level POS patterns (Shen et al., 2016).

- Baseline+SCTB: Additionally used 4,933 sentences from the newly annotated SCTB for training the Chinese analyzers.

For testing, we used the remaining 200 sentences from the newly annotated SCTB. The significance tests were performed using the bootstrapping method (Zhang et al., 2004).

Tables 1 and 2 show the word segmentation, and the joint segmentation and POS tagging results, respectively. We can see that SCTB significantly improves both segmentation and joint segmentation and POS tagging performance by a large margin, i.e., 4.27% and 7.26% F-Measure, respectively.

Table 3 shows the parsing results. We used the Evalb toolkit[7] for the parsing accuracy calculation. As Evalb was originally designed for English, it only can evaluate the sentences that have the same segmentation as the gold data. For this reason, we showed the results based on gold segmentations in Table 3. As a reference, the parsing F-Measures from scratch for Baseline and Baseline+SCTB are 74.88% and 79.80% for 66 and 107 valid sentences (sentences that have the same segmentation as the gold data), respectively.

System	Precision	Recall	F-Measure
Baseline	67.33	72.84	69.97
Baseline+SCTB	74.82	78.89	76.80†

Table 3: Parsing results based on gold segmentations ("†" indicates that the result is significantly better than "Baseline" at $p < 0.01$).

We investigated the analyses for further understanding of the improvements. Based on our investigation, we found that most of improvements come from the domain knowledge introduced by our annotated treebank. Figure 2 shows such an example. The Baseline system incorrectly segments "骨骼亚(skeleton sub)" as one word, because it lacks the knowledge that "骨骼 (skeleton)" is a medical/biology term that should be one word. This segmentation error further propagates to POS tagging and parsing. In contrast, with the help of the scientific domain knowledge introduced by our annotated treebank, the
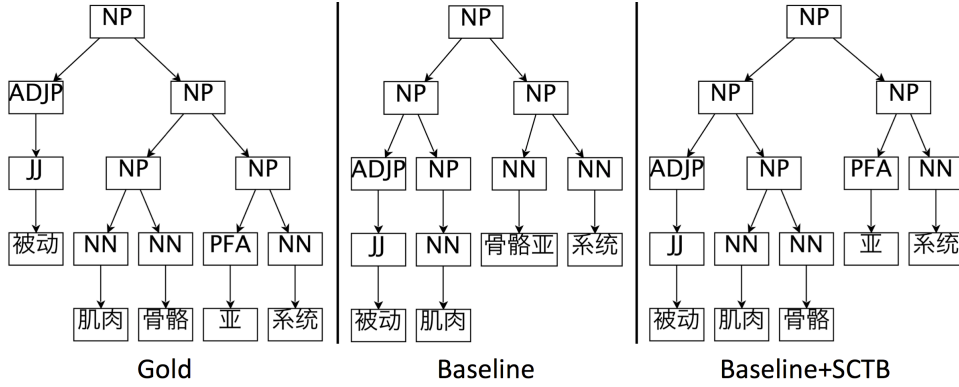
---
[6]https://github.com/slavpetrov/berkeleyparser
[7]http://nlp.cs.nyu.edu/evalb/

Figure 2: An improved example for Chinese analysis of a noun phrase "被动(passive) /肌肉 (muscle) / 骨骼 (skeleton) /亚(sub) /系统(system)".
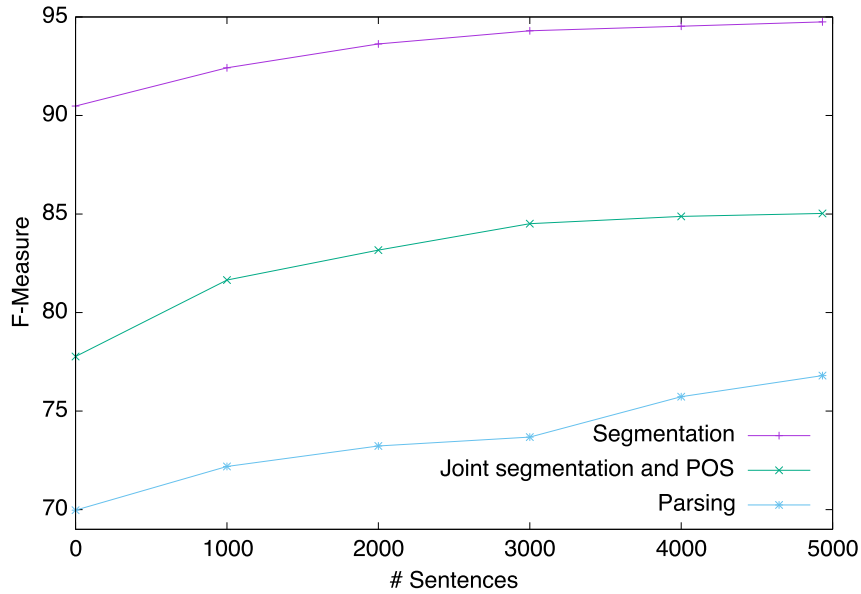

Figure 3: Chinese analysis results by adding different numbers of sentences from the newly annotated SCTB to the baseline treebanks for training the Chinese analyzers.

Baseline+SCTB system correctly segmented "骨骼 (skeleton) /亚(sub)" into two words, which also improves the POS tagging and parsing accuracy. However, the Baseline+SCTB system still fails to parse this phrase correctly, and we hope that the annotation of more sentences could be helpful for this.

To investigate the effectiveness of the treebank annotation in detail, we further conducted Chinese analysis experiments that trained the Chinese analyzers using different numbers of sentences from SCTB. In our experiments, we incrementally added 1,000 sentences to the baseline treebanks for training the analyzers. Figure 3 shows the results. We can see that for segmentation and POS tagging, the accuracy improvements slow down when more annotated sentences are used for training the analyzers; while for parsing, there is still a large potential of improvement by annotating more sentences.

3.2 MT Experiments

For Chinese-to-Japanese translation, we conducted experiments on the scientific domain MT task on the Chinese-Japanese paper excerpt corpus (ASPEC-CJ)[8] (Nakazawa et al., 2016), which is one subtask of the workshop on Asian translation (WAT)[9] (Nakazawa et al., 2015). The ASPEC-CJ task uses 672,315, 2,090, and 2,107 sentences for training, development, and testing, respectively. For Chinese-to-English translation, we conducted experiments on the Chinese-English subtask (NTCIR-CE) of the patent MT

---

[8]http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[9]http://orchid.kuee.kyoto-u.ac.jp/WAT/

63

System	ASPEC-CJ	NTCIR-CE
Baseline	39.12	33.19
Baseline+SCTB	40.08†	33.90†

Table 4: BLEU-4 scores for ASPEC-CJ and NTCIR-CE translation tasks ("†" indicates that the result is significantly better than "Baseline" at $p < 0.01$).

Source	环形碗状壁区段２６限定了开口２８，浸出液腔室 ２９ 位于 壁２４中开口２８之下。
Reference	the annular bowl wall section 26 defines an opening 28 and a leachate chamber 29 is **located** in the wall 24 beneath the opening 28.
Baseline	annular 碗状壁区 section 26 defines opening 28, leach liquid chamber 29 in the wall 24 opening 28 below.
Baseline +SCTB	the annular bowl-shaped wall section 26 defines opening 28, leach liquid chamber 29 is **positioned** below the opening 28 in the wall 24.

Table 5: An improved MT example.

task at the NTCIR-10 workshop[10] (Goto et al., 2013). The NTCIR-CE task uses 1,000,000, 2,000, and 2,000 sentences for training, development, and testing, respectively.

We used the Moses tree-to-string MT system (Koehn et al., 2007) for all of our MT experiments. In our experiments, Chinese is in the tree format, and Japanese/English is in the string format. For Chinese, we used KyotoMorph for segmentations and the Berkeley parser for joint POS tagging and parsing. We binarized the parsing results for better translation rule extraction. We compared the MT performance of the "Baseline" and "Baseline+SCTB" settings in Section 3.1. For Japanese, we used JUMAN[11] (Kurohashi et al., 1994) for the segmentation. For English, we tokenized the sentences using a script in Moses. For the Chinese-to-Japanese MT task, we trained a 5-gram language model for Japanese, on the training data of the ASPEC-CJ corpus using the KenLM toolkit[12] with interpolated Kneser-Ney discounting. For the Chinese-to-English MT task, we trained a 5-gram language model for English, on the training data of the NTCIR-CE corpus using the same method. In all of our experiments, we used the GIZA++ toolkit[13] for word alignment; tuning was performed by minimum error rate training (Och, 2003), and it was re-run for every experiment.

Table 4 shows the translation results. The significance tests were performed using the bootstrap resampling method (Koehn, 2004). We can see that the significant improvements on Chinese analysis due to the annotated treebank, also lead to the significant MT performance improvements. Despite the language pair and slight domain difference, similar improvements are observed on both the ASPEC-CJ and NTCIR-CE MT tasks.

To further understand the reasons for the improvements, we also investigated the translation results. We found that most of the improvements are due to analysis improvements of the source sentences. Table 5 shows an improved MT example from the NTCIR-CE task. We can see that there is an out-of-vocabulary word "碗状壁区 (bowl wall section)" in the Baseline result. This is because the Baseline system incorrectly segmented "碗状壁区段 (bowl wall section)" into two words "碗状壁区 (bowl wall) /段 (section)"; while the Baseline+SCTB system correctly segmented it as "碗状 (bowl) / 壁 (wall) / 区段 (section)" leading to a correct translation. Another problem of the Baseline translation is that the word "位于 (located)" is not translated. This happens because as shown in Figure 4, the Baseline system analyzed the entire Chinese phrase after the comma as a verb phrase with the word "浸出 (leach)" as the head. In contrast, although the analysis by the Baseline+SCTB system is not fully correct, it correctly analyzed the word "位于 (located)" as the head of the following verb phrase, leading to a correct translation. Both the Baseline and Baseline+SCTB systems incorrectly translated "浸出液腔室 (leachate

---

[10]http://ntcir.nii.ac.jp/PatentMT-2/
[11]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[12]https://github.com/kpu/kenlm
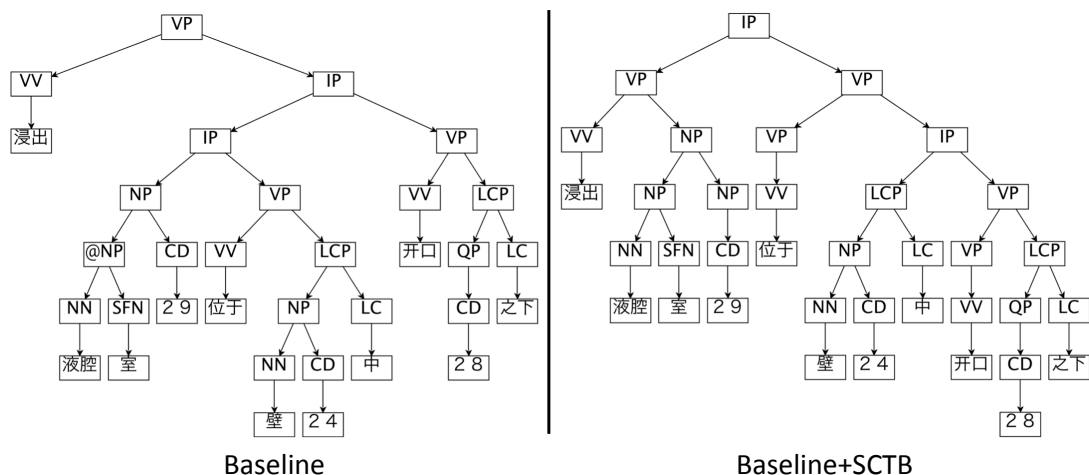[13]http://code.google.com/p/giza-pp

Figure 4: The analysis results for the Chinese phrase after the comma of the source sentence in Table 5 "浸出 (leach) /液 (liquid) /腔室 (chamber) /２９/位于 (is located) /壁 (wall) /２４/中 (in) /开口 (opening) /２８/之下 (beneath)".

chamber)" into "leach liquid chamber", this is due to the similar analysis results of both systems, while the correct analysis for this noun phrase should be "(NP (NP 浸出_NN 液_SFN) 腔室_NN) (leachate chamber)".

4 Related Work

Besides the widely used CTB (Xue et al., 2005), there are two other treebanks for Chinese. The Peking University (PKU) annotated a Chinese treebank, firstly only for segmentations and POS tags (Yu et al., 2003), and later also for syntax (Qiu et al., 2014). The Harbin Institute of Technologys (HIT) also annotated a treebank for dependency structures (Che et al., 2012). Besides the difference in annotation standards and syntactic structures, all the three treebanks are in news domain. CTB selected the raw sentences from People's Daily, Hong Kong newswire, Xinhua newswire etc., and PKU and HIT selected the raw sentences from People's Daily newswire. To the best of our knowledge, our treebank is the first publicly available Chinese treebank in scientific domain.

Three are two types of syntactic grammars for treebanking: phrase structures and dependency structures. We adopt the phrase structures used in CTB (Xue et al., 2005), because phrase structures can be converted to dependency structures based on predefined head rules using e.g. the Penn2Malt toolkit.[14] Treebanks with multi-view of both phrase structures and dependency structures also have been proposed (Qiu et al., 2014).

Recently, with more needs of multilingual NLP, the interests of constructing multilingual treebanks have increased. Multilingual treebanks such as the universal dependency treebank[15] (Nivre et al., 2016) and the Asian language treebank (Thu et al., 2016) are being constructed. As the raw sentences of our treebank were selected from parallel data and the translated Japanese and English sentences are available, we leave the potential to develop our treebank to a trilingual one.

5 Conclusion

In this paper, we presented the details of the annotation of SCTB: a Chinese treebank in scientific domain. Experiments conducted for Chinese analysis and MT verified the effectiveness of the annotated SCTB. As future work, firstly, we plan to annotate more sentences, and we aim to finish the annotation for 10k sentences within this year. Secondly, we also plan to annotate the Japanese and English raw sentences to further develop it to a trilingual treebank.

---

[14] http://stp.lingfil.uu.se/~nivre/research/Penn2Malt.html
[15] http://universaldependencies.org

Acknowledgements

References

Wanxiang Che, Zhenghua Li, and Ting Liu. 2012. Chinese dependency treebank 1.0. In *Linguistic Data Consortium*.

Huiming Duan, Xiaojing Bai, Baobao Chang, and Shiwen Yu. 2003. Chinese word segmentation at peking university. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 152–155, Sapporo, Japan, July. Association for Computational Linguistics.

Isao Goto, Ka-Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the patent machine translation task at the ntcir-10 workshop. In *Proceedings of the 10th NTCIR Conference*, pages 260–286, Tokyo, Japan, June. National Institute of Informatics (NII).

Chu-Ren Huang, Keh-Jiann Chen, and Li-Li Chang. 1996. Segmentation standard for chinese natural language processing. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 1045–1048, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28.

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, June.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 1–28, Kyoto, Japan, October.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Paris, France, May. European Language Resources Association (ELRA).

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

---

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

Slav Petrov and Dan Klein. 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*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Likun Qiu, Yue Zhang, Peng Jin, and Houfeng Wang. 2014. Multi-view chinese treebanking. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 257–268, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Ayaka Saka and Masatsura Igami. 2015. Benchmarking scientific research 2015. pages 1–172. Ministry of Education, Culture, Sports, Science and Technology, Japan, August.

Mo Shen, Hongxiao Liu, Daisuke Kawahara, and Sadao Kurohashi. 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)*, pages 253–258, Baltimore, Maryland, June. Association for Computational Linguistics.

Mo Shen, Li Wingmui, HyunJeong Choe, Chenhui Chu, Daisuke Kawahara, and Sadao Kurohashi. 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*, Osaka, Japan, December. Association for Computational Linguistics.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the asian language treebank (alt). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for chinese text annotation. In *In Proceedings of the Second Language Resources and Evaluation Conference*.

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(2):207–238, June.

Shiwen Yu, Huiming Duan, Bin Swen, and Bao-Bao Chang. 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.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, pages 2051–2054, Lisbon, Portugal, May. European Language Resources Association (ELRA).