Towards Indonesian Part-of-Speech Tagging: Corpus and Models

Sihui Fu[1], Nankai Lin[1], Gangqin Zhu[2], Shengyi Jiang[1(✉)]
[1] School of Information Science and Technology
Guangdong University of Foreign Studies, Guangzhou, China
[2] The Faculty of Asian Languages and Cultures
Guangdong University of Foreign Studies, Guangzhou, China
sihuifu93@gmail.com, neakail@outlook.com,
199210621@oamail.gdufs.edu.cn, jiangshengyi@163.com

Abstract

As a member of the Malayo-Polynesian languages, Indonesian is spoken by a large population. However, language resources and processing tools for Indonesian are quite limited. Part-of-speech (POS) tagging aims to assign a particular POS to a word, concerning its distribution and function in the context, which can provide valuable information for most natural language processing tasks. This work introduces our work on designing an Indonesian part-of-speech (POS) tagset, including 29 tags, and constructing a large Indonesian POS corpus comprised of over 355,000 tokens. During the design and annotation processes, we make judgments mostly from a typological perspective, following the specifications of Universal Dependencies, while not missing those language-specific phenomena. In addition, we try to utilize several state-of-the-art sequence labeling models, trained on the proposed corpus, to implement automatic POS tagging, and the experiment results are favorable, with the accuracies higher than 94%.

Keywords: Indonesian, part-of-speech tagging, corpus

1. Introduction

The part-of-speech (POS), also referred to as the grammatical category of a word, signifies the morphological and syntactic behaviors of a lexical item. Some common ones include verbs, nouns, adjectives and adverbs. POS tagging is the process of assigning a particular POS to a word based on both its definition and its context. Since POS can provide valuable linguistic information, POS tagging is an underlying step for most natural language processing (NLP) tasks, such as chunking, syntactic parsing, word sense disambiguation, and machine translation.

Bahasa Indonesia (Indonesian for 'language of Indonesia') is a member of the Malayo-Polynesian branch of the Austronesian language family. Unlike English or other high-resource languages, although spoken by over 198 million people (Simons and Fennig, 2017), Indonesian possesses quite limited language resources, which also leads to the limited development of language technology applied to Indonesian.

Some previous studies have presented their efforts on the construction of Indonesian POS corpora or automatic POS taggers (Dinakaramani et al., 2014; Pisceldo et al., 2009; Rashel et al., 2014), but to the best of our knowledge, either the size of the corpora they used is not big, or the taggers do not attain satisfactory performance.

In this paper, we report our work on designing an Indonesian POS tagset and building a large manually-tagged Indonesian POS corpus comprised of over 355,000 tokens, under the instructions of Universal Dependencies[1]. Furthermore, we attempt to achieve automatic POS tagging using state-of-the-art models.

In the remaining parts, section 2 will briefly review previous work on Indonesian POS tagging. The design and construction processes are described in section 3. The models we employ to build POS taggers are introduced in section 4. Section 5 gives experiment setups and results and section 6 concludes.

2. Related Work

In terms of Indonesian POS tagging, only few corpora are available and relevant processing tools are not mature enough. Pisceldo et al. (2009) employed two POS tag schemes (containing 37 and 25 tags respectively) to manually annotate two Indonesian corpora[2], and intended to develop an Indonesian POS tagger based on conditional random fields (CRF) and maximum entropy (ME). However, the size of their corpora is small (40,513 tokens in total). Also, the experiment results on corpus 1 are not ideal (The highest accuracy is 77.36% for 37 tags and 85.02% for 25 tags).

Wicaksono and Purwarianti (2010) built a Hidden Markov Model (HMM) based POS tagger on a 15,000-token Indonesian corpus, which was proposed in Pisceldo et al. (2009). Affix tree, succeeding POS tags and additional dictionary lexicon were used to improve the performance of vanilla HMM. Subsequently, they extended their work to develop an Indonesian Mind Map Generator (Purwarianti et al., 2013), which includes several Indonesian NLP tools such as POS tagger, syntactic parser and semantic analyzer. The POS tagger was built based on the methods mentioned in their 2010 work, while a decision tree was also used to handle the empty score of emission probability.

Dinakaramani et al. (2014) explored the design of a linguistically motivated Indonesian POS scheme,

[1] http://universaldependencies.org/
[2] For each corpus, they tried both POS tag schemes.

and manually tagged a corpus containing 10,000 sentences (over 250,000 tokens). Furthermore, they developed a rule-based POS tagger by combining several language resources, including closed-class tagging dictionary, multi-word expression dictionary, MorphInd (Larasati et al., 2011) and disambiguation rules, and then applied this tagger to their previously proposed corpus, obtaining an accuracy of 79% (Rashel et al., 2014).

In this work, we propose our own Indonesian POS tagset and present a larger Indonesian POS corpus, compared with previous work. Moreover, we attempt to build an automatic POS tagger based on state-of-the-art models.

3. The Corpus

3.1 The Design of Indonesian Tagset

Both Indonesia and English employ the Latin alphabet as their writing system and (almost) contain the same letters. On the other hand, many Indonesian words are derived from English words, such as *komputer* 'computer', *halo* 'hello', *mesin* 'machine', etc. Given these similarities, we first based our design of Indonesian POS tagset on the existing English ones, and chose the Penn Treebank tagset (Santorini, 1990) for its maturity and popularity. Furthermore, by virtue of the guidelines of Universal Dependencies, we regulated our initial tagset to achieve cross-linguistically consistent annotation, while attending to language-specific phenomena. In addition, during the manual annotation, we also consulted previous work on Indonesian tagsets (Dinakaramani et al., 2014; Pisceldo et al., 2009) to see if any revision was required.

One of the guiding principles is simplicity. Different corpora adopt different tagging schemes, which leads to varying sizes of tagset. For example, there are 87 tags in the Brown Corpus tagset, 45 in the English Penn Treebank tagset, whereas 137 tags in the UCREL CLAWS7 tagset. Considering that manual annotation of a large scale corpus is labor-intensive, a tagset consisting of lots of tags will increase annotators' cognitive load, and therefore we should propose a small tagset, while maintaining useful linguistic information for later natural language processing tools. Meanwhile, we want to develop a corpus which can describe the common properties and structural diversities of multiple languages, from the perspective of linguistic typology. Hence, following the instruction of Universal Dependencies, we abstracted those widespread grammatical categories found cross-linguistically (universality), but did not ignore those specific ones in Indonesian (particularity).

3.2 Data Source

To obtain attested Indonesian data, we crawled substantive news articles from the website detik.com [3], Indo-Asia-Pacific Defense Forum [4], BBC Indonesia[5], etc., whose content covers various topics including politics, finance, society, military, etc. After separating paragraphs into individual sentences, we randomly picked out over 20,000 sentences as our dataset to be annotated. Altogether, the corpus has over 355,000 lexical tokens.

3.3 The Annotation Process

This section will briefly introduce the annotation process of our dataset. In general, the process contains five steps and seven human annotators are engaged in.

(1) The first 2000 sentences are manually annotated, according to the initial tagset constructed on the basis of the Penn Treebank tagset. Annotators need to annotate each word in a sentence by means of its syntactic function and definition in the KBBI dictionary [5]. In this step, considerable issues are put forward, such as the adequate tags for abbreviations, combinations of digits and letters, book titles, and website links. Solutions are presented after discussions and agreement of all annotators and the tagset was revised accordingly.

(2) Referring to the specifications in Universal Dependency, we further regulated our tagset. Thus, the definition of a grammatical category in Indonesian is more consistent with that in other languages. However, specific properties could not be neglected, such as the pronominal suffix in the preposition-object structure, *olehnya* 'by him/her'. Therefore, several language-specific POS tags are also proposed.

(3) The first 2000 sentences were retagged. Annotators should make their judgments based on the specifications in Universal Dependencies and syntactic information. Issues were welcomed to raise and solved by joint discussions.

(4) The remaining sentences were manually tagged, in accordance with the procedure described in (3).

(5) We manually evaluated and revised the tagged sentences with the help of annotators. At the same time, some previous work was reviewed to make comparisons and revisions.

3.4 Problematic Cases

Unlike English, it seems that Indonesian does not have a standardized grammar system by far, which has brought about plentiful confusions and disputes to our design of POS tags and the process of data annotation. On the one hand, different people have different opinions on a word as to its grammatical category. For instance, *sudah* 'already' is regarded as an adverb in KBBI, but as a modal verb in Dinakaramani et al. (2014); *sekarang* 'now' is regarded as an adverb in Pisceldo et al. (2009) but a common noun in Dinakaramani et al. (2014). According to Tallerman (2015), three important linguistic criteria for identifying a word's class is to check its morphosyntax, distribution and function in

[3] https://news.detik.com/

[4] http://apdf-magazine.com/id/

[5] http://www.bbc.com/indonesia

a phrase or sentence. However, Indonesian is not an inflectional language, which means morphosyntax may not help. Therefore, though annotators would rely more on Indonesian dictionaries at the beginning, in the subsequent stages we required them to make judgments based on a word's distribution and function in its context, instead of being bound to the POS given by dictionaries.

On the other hand, it is difficult to achieve the balance between universality and particularity. In Indonesian grammar, there exist some special grammatical categories, which we might find their alternatives in the universal framework. In such case, whether to retain the original terms or to incorporate them in the unified framework is a problem, since rough incorporation may lose the traits of the individual language. A typical example is those indefinite numbers in Indonesian, including *beberapa* 'some', *semua* 'all', *banyak* 'many', etc., which may be regarded as indefinite pronouns in English. However, we noticed other indefinite numbers like *belasan* 'eleven to nineteen' and *ratusan* 'hundreds' share more similarities with numbers. Thus, we at last preserve the category 'indefinite number'.

In addition, since Indonesian is an agglutinative language, many of its complex words are formed by stringing together multiple morphemes (including stems and affixes) without changing their spellings. One case is those words with pronominal suffixes, such as *namamu* 'your name' (*nama* 'name', *-mu* 'your'), *olehnya* 'by him/her/them' (*oleh* 'by', *-nya* 'him/her/them'), etc. Figure 1 lists three cases concerning the use of the pronominal suffix *–nya* ('INJ' suggests interjection). Some previous work separates such words into the stems and suffixes and tags them respectively. However, we insist to maintain a word's integrity, and therefore propose three unique tags: SP (subject-predicate relation), VO (verb-object relation) and PO (preposition-object relation), corresponding to the three cases in Figure 1 respectively. One might argue that such words should be tagged according to the grammatical categories of the heads in these structures. We will not deal with it in this work, leaving it open for discussions.

(1) "Tapi saya tidak marah kok", **katanya**.
But I not angry INJ said.she
"But I am not angry!", she said.
(2) Orang tuanya **mengusirnya** dari rumah.
People old.his expelled.him from home
His parents threw him out of the house.
(3) Mesin yang rusak ini diperbaiki **olehnya**.
machine that broken this repaired by.him
The broken machine was repaired by him.

Figure 1. Several cases of the pronominal suffixes *–nya* in Indonesian

3.4 Indonesian POS tagset

The final version of our Indonesian POS tagset is presented in Table 1, consisting of 29 tags.

Tag	Description	Example
CC	Coordinating conjunction	dan, tetapi, atau
CD	Cardinal number	satu, dua, tiga, 79, 2017, 0.1
DT	Determiner	para, sang, si, sebuah, seorang
FW	Foreign word	poetry, technology, out, world
ID	Indefinite number	puluhan, segala, 30-an, beberapa
IN	Preposition	di, ke, oleh, untuk, dari, antara
JJ	Adjective	besar, tinggi, manis, cerdik
JJS	Adjective, superlative degree	terdekat, terbesar, terpenting, terbaik
MD	Auxiliary verb	harus, perlu, boleh, adalah, mau
NN	Common noun	buku, pipi, rupiah, km, sekarang
NNP	Proper noun	Indonesia, MH370, Li Li, SBY
OD	Ordinal number	pertama, ketiga, ke-6
P	Particle	pun, -lah, -kah
PO	Preposition-object structure	untuknya, antaranya, olehku, padamu
PRD	Demonstrative pronoun	ini, itu, sini, sana
PRF	Reflexive pronoun	sendiri, diri
PRI	Indefinite pronoun	siapapun, apapun
PRL	Relative pronoun	yang
PRP	Personal pronoun	saya, kamu, dia, kami, kalian
RB	Adverb	sudah, tidak, sangat, juga
SC	Subordinating conjunction	baik...maupun..., sebelum, kalau
SP	Subject-predicate structure	katanya, sebutnya, tuturnya, imbuhnya
SYM	Symbol	+, %, @, $, 15/2/2017, 13:00
UH	Interjection	oh, hai, ya, sih, mari
VB	Verb	ada, melihat, gagal, menyoroti, main
VO	Verb-object structure	meningkatnya, terbentuknya
WH	Question	apa, siapa, mana, bagaimana
X	Unknown	yagg, busaway, saaat
Z	Punctuation	",.?"()

Table 1. Indonesian POS tagset

4. Models

POS tagging can be regarded as a sequence labeling problem. Given an input sequence $X = \{x_1, x_2, \dots, x_n\}$, where n is the length of X, the prediction model should output a sequence $Y = \{y_1, y_2, \dots, y_n\}$, in which each y_i is the label of x_i. In the case of POS tagging, Y refers to the POS sequence of an input sentence. State-of-the-art supervised models for handling the sequence labeling problem include conditional random fields (CRFs) (Lafferty et al., 2001), long short-term

memory (LSTM) (Huang et al., 2015; Lample et al., 2016; Reimers and Gurevych, 2017), etc. In this paper, we explore three models to achieve automatic POS tagging, namely CRFs, Bidirectional LSTM (Bi-LSTM) and sequence-to-sequence learning (seq2seq).

4.1 CRFs

CRFs are a type of discriminative undirected graphical model for labeling sequential data. For a linear chain CRF, given an input sentence s, the score of one of its possible label sequence l can be calculated through Equation 1:

$$sc(l|s) = \sum_{j}^{m} \sum_{i}^{n} \lambda_j f_j(l_{i-1}, l_i, i, s) \quad (1)$$

where i is the position of a word in the sentence, l_i is the label of the current word, l_{i-1} is the label of the previous word, f_j is the feature function, and λ_j is the feature weight. After having the scores of each possible label sequence, we can obtain the probabilities of these label sequences by exponentiation and normalization:

$$p(l|s) = \frac{\exp\left[sc(l|s)\right]}{\sum_{l'} \exp\left[sc(l'|s)\right]} = \frac{1}{Z(s)} \exp\left[sc(l|s)\right] \quad (2)$$

where $Z(s)$ is usually called the normalization factor.

4.2 Bi-LSTM

LSTM networks have been widely used in many sequence labeling tasks and show state-of-the-art performance. In this work, we employ the Bi-LSTM network with a CRF classifier (Fig.2, slightly adapted from (Reimers and Gurevych, 2017)). Its character representation is also derived from a Bi-LSTM network (Fig.3). A detailed explanation of the Bi-LSTM model can be found in (Huang et al., 2015; Lample et al., 2016).

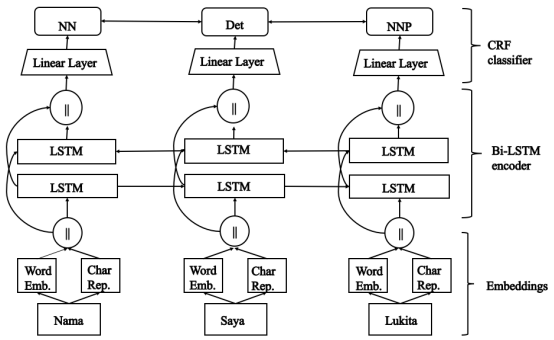

Figure 2. The Bi-LSTM network with a CRF classifier. (||) means concatenation.

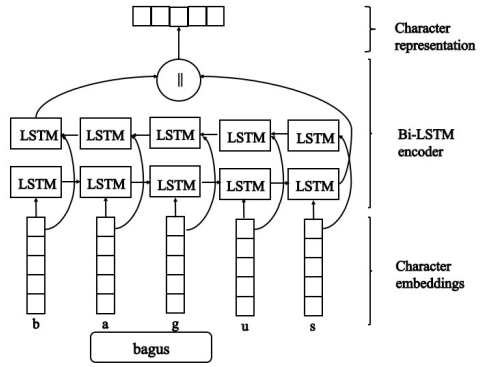

Fig.3 Character-based representation derived from the Bi-LSTM network. (||) means concatenation.

4.3 Seq2seq

Sequence to sequence learning has been successfully applied to machine translation (Wu et al., 2016) and text summarization (Nallapati et al., 2016). A popular approach is to encode an input sequence into a distributed representation with a bi-directional recurrent neural network (RNN) and decode the representation with another RNN, while the encoder and decoder are usually linked by the attention mechanism (Ghader and Monz, 2017), as Figure 4 shows. In this work, we also attempt to use the sequence-to-sequence architecture to perform sequence labeling.

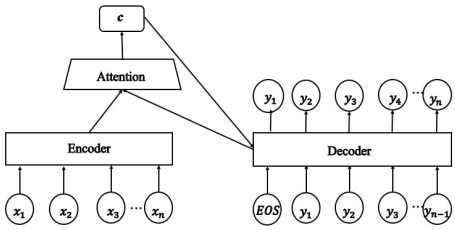

Figure 4. The sequence-to-sequence framework

5. Experiment

5.1 Setup

For CRFs, we used CRF++[6], a simple and open-source implementation of CRFs. Table 2 lists the feature set which obtained the best performance in our experiments, and we report the experiment result based on this feature set. Except for the *-c* (which was set as 3), other parameters are in accordance with the default settings.

Type	Feature	Description
Unigram	w_n $(n = -1,0,1)$	The previous n, current, and next n words
Prefix	$p_n(w_0)$, $n = 2,3,4$	The first n letters in the current word
Suffix	$s_n(w_0)$, $n = 2,3,4$	The last n letters in the current word
Bigram	$t(w_{-1})$	The predicted tag of the previous word

Table 2. The defined feature sets used in CRFs

[6] http://taku910.github.io/crfpp/

To implement the Bi-LSTM network, we used TensorFlow[7] version 1.2. For the setting of hyper-parameters, we referred to the suggestions of Reimers and Gurevych (2017). The pre-trained word embedding, with 200 dimensions, was trained on the Indonesian text (about 170 million tokens) that was crawled from several Indonesian news websites, using GloVe (Pennington et al., 2014). If a token does not occur in the vocabulary of the pre-trained word embedding, we would assign it a random word embedding (subject to a Gaussian distribution). In addition, the pre-trained word embedding is trainable during the training process. The dimension of character embedding is 100. The number of recurrent units for the Bi-LSTM layer which produces character representation (Fig.3) is 100, while for another Bi-LSTM layer (Fig.2) is 300. Adam was chosen as the optimizer. The dropout rate is 0.5. Also, we used a mini-batch size of 32 and employed the early stopping strategy if the score for development set does not increase for more than 3 training epochs. We report the result from the run with the highest score on development set.

As for the seq2seq architecture, we used NeuralMonkey [8], a convenient tool for quickly building sequential neural network models. It has implemented a framework for tagging. Therefore, the SentenceEncoder module was employed as the encoder, and the SequenceLabeler module as the decoder. Hyper-parameters are in accordance with the default settings.

5.2 Result

To compare the performance of different models, we employed the 10-fold cross validation. The corpus was divided into 10 folds. In each experiment, we used one fold for test and the remaining for training. Accuracies of different models can be calculated by comparing the manual tagging and the automatic tagging realized by these models. Table 3 shows the results. For each model, we report the average accuracy of 10 experiments.

Models	Avg Acc.
CRFs	95.12%
Bi-LSTM+CRF	95.68%
Seq2seq	94.14%

Table 3. The performance of different models

The highest average accuracy is produced by the Bi-LSTM network with a CRF classifier. CRFs perform slightly worse. It seems that Seq2seq is not competitive with the other two methods, but it takes the least time to train the model. Next, we will consider utilizing different encoders and decoders, and adding the attention mechanism to improve the performance of the Seq2seq architecture.

6. Conclusion

This paper describes our work on designing an Indonesian POS scheme and building a considerable Indonesian POS corpus using the text collected from multiple sources. In the design process, it is important to make a trade-off between universality and particularity. We put emphasis on those grammatical categories found cross-linguistically, following the specifications of Universal Dependencies, but would not miss those specific ones in Indonesian. During the annotation process, to tag a word, annotators need to consider its distribution and function in the context, not only the POS given by dictionaries. Finally, we propose an Indonesian POS tagset comprised of 29 tags and an Indonesian POS corpus of over 355,000 tokens, which could contribute to Indonesian language resources and provide support for further Indonesian NLP. Furthermore, we tried to achieve automatic POS tagging by using several state-of-the-art models trained on our corpus, and the experiment results are quite promising.

In future work, we intend to build a high-performance Indonesian POS tagger. Moreover, we would like to use the corpus to aid other Indonesian NLP tasks, such as chunking, syntactic parsing, etc.

Acknowledgements

This research was substantially supported by the National Natural Science Foundation of China (No. 61572145) and Department of Education of Guangdong Province. We greatly thank our annotators for their excellent wok.

References

Dinakaramani, A., Rashel, F., Luthfi, A., and Manurung, R. (2014). Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014* (pp. 66–69).

Ghader, H., and Monz, C. (2017). What does Attention in Neural Machine Translation Pay Attention to? *arXiv preprint arXiv: 1710.03348*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv: 1508.01991*.

Kamus Besar Bahasa Indonesia. (2008). Kamus Pusat Bahasa, Jakarta, 4th edition.

Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (Vol. 8, pp. 282–289).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL 2016*. San Diego, California.

Larasati, S. D., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Communications in Computer and Information Science* (Vol. 100 CCIS, pp. 119–129).

[7] https://www.tensorflow.org/

[8] https://github.com/ufal/neuralmonkey

Nallapati, R., Zhou, B., Santos, C. N. dos, Gulcehre, C., and Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Pisceldo, F., Adriani, M., and Manurung, R. (2009). Probabilistic Part of Speech Tagging for Bahasa Indonesia. In *Proceedings of the 3rd International MALINDO Workshop, Colocated Event ACL-IJCNLP*.

Purwarianti, A., Saelan, A., Afif, I., Ferdian, F., and Wicaksono, A. F. (2013). Natural language understanding tools with low language resource in building automatic indonesian mind map generator. *International Journal on Electrical Engineering and Informatics*, *5*(3), 256–269.

Rashel, F., Luthfi, A., Dinakaramani, A., and Manurung, R. (2014). Building an Indonesian rule-based part-of-speech tagger. In *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014* (pp. 70–73).

Reimers, N., and Gurevych, I. (2017). Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. *arXiv preprint arXiv: 1707.06799*.

Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). http://doi.org/10.1017/CBO9781107415324.004.

Simons, G. F., and Fennig, C. D. (Eds.). (2017). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 20th edition.

Tallerman, M. (2015). Understanding Syntax. Routledge, Taylor & Francis Group, London, 4th edition.

Wicaksono, A. F., and Purwarianti, A. (2010). HMM Based Part-of-Speech Tagger for Bahasa Indonesia. In *4th International MALINDO (Malay and Indonesian Language) Workshop*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv: 1609.08144*.