An Overview of BPPT's Indonesian Language Resources

Gunarso, Hammam Riza, Elvira Nurfadhilah,
M. Teduh Uliniansyah, Agung Santosa, Lyla R. Aini

Agency for the Assessment and Application of Technology (BPPT), Jakarta, INDONESIA

Abstract

This paper describes various Indonesian language resources that Agency for the Assessment and Application of Technology (BPPT) has developed and collected since mid 80's when we joined MMTS (Multilingual Machine Translation System), an international project coordinated by CICC-Japan to develop a machine translation system for five Asian languages (Bahasa Indonesia, Malay, Thai, Japanese, and Chinese). Since then, we have been actively doing many types of research in the field of statistical machine translation, speech recognition, and speech synthesis which requires many text and speech corpus. Most recent cooperation within ASEAN-IVO is the development of Indonesian ALT (Asian Language Treebank) has added new NLP tools.

1 Introduction

As a national language of Indonesia, Bahasa Indonesia has been used as a lingua franca in the multilingual Indonesian archipelago for centuries. Indonesia is the fourth most populous nation in the world, after China, India and the United States. Of its large population, around 255 million people, the majority speak Indonesian, making it one of the most widely spoken languages in the world.

Aside from speaking the national language, most Indonesians are fluent in any of more than 746 distinct regional languages (Amalia, 2016) such as Javanese, Sundanese and Madurese, which are commonly used at home and within the local community. Most formal education, and nearly all national media and other forms of communication, are conducted in Indonesian. Throughout the archipelago, Bahasa Indonesia has become the language that bridges the language barrier among Indonesians who have different mother-tongues.

In recent years, countries in the same region tend to establish some free trade areas such as ASEAN Economic Community (AEC), European Union (EU), and Asia-Pacific Economic Cooperation (APEC). This opens opportunities to accelerate economic growth for Indonesia. However, these efforts are hindered due to the lack ability of Indonesians in communicating with foreigners.

BPPT has started collecting language resources since 1987 as part of the development of a multilingual machine translation system in a project called "The Research and Development Cooperation Project on a Machine Translation System for Japan and its Neighbouring Countries". At the end of the project, many Indonesian language resources have been resulted, such as Indonesian basic dictionary, Indonesian grammar rule for analysis and generation system, Indonesian monolingual text corpus, and Indonesian gazetteer.

We have continued collecting language resources to improve the system which has been developed and the development of other natural language processing systems. The needs for the development of statistical machine translation with Indonesian as source language, Indonesian speech recognition and Indonesian speech synthesizer led to the development of other language resources, which are parallel corpora and speech corpora for ASR and TTS.

2 Indonesian Gazetteer

Indonesia is the world's largest island country, with more than thirteen thousand islands and has 34 provinces, of which five have Special Administrative status. Indonesia consists of hundreds of distinct

This research is funded by government research budget for BPPT fiscal year 2016 and cooperation with NICT ASEAN-IVO

Proceedings of the 12th Workshop on Asian Language Resources,
pages 73–77, Osaka, Japan, December 12 2016.

native ethnic and linguistic groups. This big country also has many districts, regencies, lakes, mounts, ports, airports, rivers, capes, bays, etc. Regarding this diversity, it needed to compile Indonesian Gazetteer as one of the language resources. Table 1 lists entries of the Indonesian Gazetteer.

No	Name Entity	Number of Data
1	Province	34
2	Regency	484
3	District	6,793
4	Lake	101
5	Mount	546
6	Airport	137
7	Harbour	295
8	Island	948
9	River	586
10	Cape	627
11	Bay	301
12	Tribe	358
13	Weapon	272
14	Art	245

Table 1: Lists entries of the Indonesian gazetteer

3 Indonesian Monolingual Corpus

Up to now, we have collected around 10.5 million sentences in an Indonesian monolingual corpus. The sentences were taken from various sources available on the internet such as national newspapers/magazines and governmental institutions (presidential speech, meeting transcriptions, trial transcriptions, etc.) by using HTTrack, a free offline browser utility (Roche et al., 2007). Table 2 lists all the corpora obtained from various sources.

Topic	Source	Number of articles	Number of sentences	Number of unique sentences	Number of words	Number of unique words
Financial	Bank of Indonesia	124	115,431	113,615	3,081,380	28,421
Various topics	DPR (House of Representative)	355	205,405	202,816	4,293,868	48,525
Law	PN (District Court)	12	39,075	38,733	662,964	17,803
Various topics	Presidential speech	16	1,268	1,266	24,695	3,502
Financial	Ministry of Finance	46	6,172	6,153	135,981	8,945
Various topics	Mail archive	3,685	68,455	56,267	1,092,195	45,323
Financial	BPK (Supreme Audit Board)	501	862,542	831,334	35,521,560	127,108
Various topics	DPD (House of Regional Representative)	755	450,270	444,836	9,902,733	72,147
Politics	KPU (National Election Commission)	1,176	23,503	16,734	399,182	19,042

Topic	Source	Number of articles	Number of sentences	Number of unique sentences	Number of words	Number of unique words
Law	Ministry of Justice and Human Rights	6,222	361,140	349,630	8,796,144	51,326
Literature	Novels	110,943	5,760,141	5,684,129	72,605,688	396,736
Various topics	National newspa-per/magazine	28,795	609,728	609,275	12,484,728	111,574
Law	MK (Constitutional Court)	7,293	1,992,251	1,912,706	36,741,176	163,397
Various topics	Combination of all above	159,923	10,495,381	10,445,098	185,602,460	647,982

Table 2: Indonesian monolingual text corpus

4 Indonesian-English Parallel Corpus

No	Source	Topic	Number of sentences	Number of unique sentences
1	ASEAN MT[1]	Tourism	21,969	19,359
2	BBC	News	5,284	5,083
3	BTEC	Tourism	133,453	127,815
4	Indonesian Ministry of Finance [2]	Economics	48,778	46,400
5	PanL [3]	Economics	6,708	6,677
6	PanL	Science and Technology	10,431	10,404
7	PanL	National news	10,141	10,141
8	PanL	Sports	14,217	14,216
9	PanL	International news	9,993	9,993
10	Tatoeba[4]	Various topics	4,179	3,694
11	U-STAR[5]	Tourism	6,500	6,451
12	Warisan Indonesia	Tourism	7,517	7,161
13	Colours Magazine Garu-da [6]	Tourism	10,603	10,400
14	Various expatriate blogs	Culture	33,943	33,943
15	Asian Language Treebank (ALT)[7]	WikiNews	20,000	20,000

Table 3: Indonesian-English parallel corpus

[1] (AseanMT, 2014)
[2] (Kemenkeu, 2015)
[3] (PanL, 2010)
[4] (Tatoeba, 2012)
[5] (Ustar, 2013)
[6] (Garuda-Indonesia, 2013)
[7] (Asian Language Treebank, ASEAN-IVO)

PTIK-BPPT collected around 311,737 sentences in an Indonesian-English parallel corpus to aid our research in statistical machine translation. The sentences were taken from various sources available on the internet such as national newspapers/magazines and governmental institutions by using HTTrack, a free offline browser utility. We hired some professional translators to check the correctness of the parallel corpus manually. Table 3 lists all parallel corpus obtained from various sources.

The Asian Language Treebank (ALT) project aims to advance the state-of-the-art Asian natural language processing (NLP) techniques through the open collaboration for developing and using ALT. The project is a joint effort of six institutes for making a parallel treebank for seven languages: English, Indonesian, Japanese, Khmer, Malay, Myanmar, and Vietnamese. In creating Indonesian - ALT, it requires tools to speed up the development. Some of these tools have been provided by the ALT project but for Indonesian we will use tools that were created from previous projects. Among them are POS Tagger, Syntax Tree Generator, Shallow Parser, word alignment, etc. Indonesian treebank resulted from this project will be utilize to enhance the exising tools and to create new tools in the field of NLP using state of the art techniques. Indonesian treebank is also expected to help the advancement of NLP researches in Indonesia.

5 Automatic Speech Recognition

To develop automatic speech recognition (ASR) system, training data in the form of speech corpus is required. The speech corpus for ASR must have a rich combination of uttered phonemes in the targeted language. And to make the ASR system speaker independent, the corpus should be created from speech recordings of many speakers with various ages and gender. Currently we have two set of speech corpus created for Indonesian ASR. The first one was made in 2010 for the joint development project with PT. INTI[7] to develop an Indonesian ASR system called PERISALAH. This speech corpus, consists of total 100,000 utterances uttered by 400 people. The utterances were coming from around 7800 unique sentences. The speakers consists of 200 adults male and 200 adults female, with the following composition: 40% Javanese, 20% Sundanese, 20% from Batak, 5% from Minang, 5% from Makassar, 5% from Maluku, and 5% from Papua, Bali and Madura. The ages of the speakers are within 20 to 50 years old. The total duration of the speech data is more than 133 hours. The average time per utterance is around 5 seconds, the longest utterance time is 22 seconds, and the shortest utterance time is 1.5 seconds. The speech data in this corpus was recorded as a single channel data with a 16KHz sampling rate and a 16-bit data size. The file format used for storing the data is WAV format. This first corpus set was already tested to create an acoustic model for Indonesian ASR with WER of around 20% using Julius[8] as the ASR engine. Since the PERISALAH corpus was created as a joint development, the ownership was a shared one, so it is not publicly available.

The second set of the corpus was created in 2013, this second corpus was planned to be made publicly available for research and education purposes. The second corpus consists of total 49,000 utterances uttered by 200 people, where each person speaks around 245 sentences. The speakers were consists of 100 male and 100 female, and the age-range were within 15 to 50 years old. Around 30% of the speakers were high-school students. The sentence used in this corpus comes from 5,000 unique sentences. The file format used to store the data is WAV with single channel recording, a 16KHz sampling rate and a 16-bit data size. The total duration of the speech data is more than 95 hours. The average time per utterance is around 6.8 seconds, the longest utterance time is 29 seconds, and the shortest utterance time is 1.0 seconds. This second corpus is planned to be released for research and education community at the end of 2016.

6 Speech Corpus for The Development of An Indonesian TTS System

For developing an Indonesian TTS system, PTIK-BPPT now uses 3 sets of speech corpus. Set 1 and 2 consists of 5,000 WAV files each, and set 3 consists of 15,645 WAV files. Table 4 describes details of the speech corpus.

Set no.	Speaker	Number of Utterances	Length (hours)	Format
1	male adult	5,000	6.56	wav, 16-bit, 16KHz
2	female adult	5,000	7.13	wav, 16-bit, 16KHz
3	male adult	15,645	40.25	wav, 16-bit, 16KHz
3	female adult	15,645	30.52	wav, 16-bit, 16KHz

Table 4: Speech TTS corpus.

The following table lists types of sentences in each speech corpus set:

Set no.	Sentence type		Sentence type			
	Number of regular sentences	Number of conversation-al sentences	Number of declarative sentences	Number of interroga-tive sen-tences	Number of imperative sentences	Number of exclamatory sentences
1	3,715	1,285	4,532	353	9	106
2	3,809	1,191	4,566	321	9	104
3	14,173	1,473	14,554	415	45	631

Table 5: Sentence types

7 Conclusion

BPPT has collected language resources to develop Indonesian-English statistical machine translation system, Indonesian ASR and text-to-speech system. The language resources will help the advancement of MT, ASR, and TTS research in Bahasa Indonesia and any NLP-related research in general. The existing data is enough for developing MT, ASR and TTS for the Bahasa Indonesia language but it needs more efforts. We have developed MT, ASR and TTS systems based on this data with adequate performance. Currently we also involved in the development of Asian Language Treebank to enrich our NLP resources. Currently all resources are for internal use only, but in the end of 2016 we are planning to release most of these resources for research and education communities under Common Criteria (CC-BY).

Reference

1. Amalia, Dora. 2016. *Lexicographic Challenges in Minority Languages of Indonesia*. Retrieved August 03, 2016 from http://compling.hss.ntu.edu.sg/events/2016-ws-wn-bahasa/pdfx/amalia.pdf

2. Asean MT. 2014. Retrieved August 03, 2016 from http://www.aseanmt.org/index.php?q=index/download

3. Garuda Indonesia. 2013. Retrieved January 15, 2014 from https://www.garuda-indonesia.com/id/en/garuda-indonesia-experience/in-flight/in-flight-entertainment/

4. PAN Localization. 2010. Retrieved September 23, 2010 from http://www.panl10n.net/

5. Roche, Xavier et al. 2007. *HTTrack Website Copier*. Retrieved May 02, 2010 from http://www.httrack.com/

6. Tatoeba. 2012. Retrieved March 15, 2012 from https://tatoeba.org/eng/

7. I-Perisalah , Retrieved August 15, 2016 from http://www.inti.co.id/index.php/id/2015-06-18-06-16-48/infrastruktur/smart-meeting

8. Lee, Akinobu, and Tatsuya Kawahara. 2009. *Recent Development of Open-source Speech Recognition Engine Julius*. Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee.