Building and Annotating the Linguistically Diverse NTU-MC (NTU — Multilingual Corpus)

Liling Tan and Francis Bond
Division of Linguistics and Multilingual Studies,
Nanyang Technological University
14 Nanyang Drive, Singapore, 637332, Singapore
alvations@gmail.com, bond@ieee.org

Submitted on 14 May, 2012

Abstract
The NTU-MC compilation taps on the linguistic diversity of multilingual texts available within Singapore. The current version of NTU-MC contains 595,000 words (26,000 sentences) in 7 languages (Arabic, Chinese, English, Indonesian, Japanese, Korean and Vietnamese) from 7 language families (Afro-Asiatic, Sino-Tibetan, Indo-European, Austronesian, Japonic, Korean as a language isolate and Austro-Asiatic). The NTU-MC is annotated with a layer of monolingual annotation (POS ans sense tags) and cross-lingual annotation (sentence-level alignments). The diverse language data and cross-lingual annotations provide valuable information on linguistic diversity for traditional linguistic research as well as natural language processing tasks. This paper describes the corpus compilation process ith the evaluation of the monolingual and cross-lingual annotations of the corpus data. The corpus is available under the Creative Commons – Attribute 3.0 Unported license (CC BY).

Keywords
Multilingual, Corpus, Annotations, Parallel texts, POS tagging, Alignments

1 Introduction

“The rapidly growing gap between the demand for high-quality multilingual content and the lag in the supply of language professionals is driving the requirement for technology that can dramatically improve translation turnaround time while maintaining exceptionally high output quality" (McCallum, 2011). Cross-lingual training using parallel corpora has been gaining popularity in NLP application tasks such as word sense disambiguation (e.g. Sarrafzadeh et al. 2011; Saravanan et al. 2010; Mitamura et al. 2007), information retrieval and question answering. In addition, parallel corpora are valuable resources for advancing linguistic annotations morphologically, syntactically and semantically (e.g. Snyder and Barzilay 2008; Hwa et al. 2005; Resnik 2004).

The essential knowledge resource for building these language technologies is parallel corpora. The present pool of resources holds a sizable amount of European parallel corpora (e.g. Ralf et al. 2006; Erjavec 2004), an increasing interest in building Asian languages-

English bitexts (e.g. Xiao et al. 2004) but only a handful of parallel Asian language corpora (e.g. Zhang et al. 2005).

To fill the lack of parallel corpora of Asian languages, the NTU–Multilingual Corpus (NTUMC) taps on the array of multilingual texts available in Singapore. Singapore's multicultural and multilingual society means that information in various languages is often found on signboards, public announcements and in widely disseminated information dissemination. The NTU-MC presents multilingual data from a modern cosmopolitan city where people interact in different languages. Empirically, the NTU-MC represents unique societal linguistic diversity; computationally, the NTU-MC provides diverse parallel text for NLP tasks. This paper discusses the compilation of the NTU-MC from data collection to the present state of POS tagged sentence-aligned parallel texts with some sense annotation.

The rest of the paper is structured as follows: Section 2 describes the sub-tasks in the corpus compilation, the monolingual annotation and cross-lingual annotation process; Section 3 present the NTU-MC outputs and evaluates the layers of annotations; Section 4 presents the work in progress on the NTU-MC and Section 5 concludes.

2 Corpus Construction

The NTU Multilingual Corpus adopts an opportunistic data collection approach looking for existing multlingual data which can be freely redistributed. Singapore has a wealth of such data. The corpus project was granted the permission to use the websites that are published by the Singapore Tourism Board (STB). We collected two domains from the STB websites: general tourism and medical tourism. The `yoursingapore` subcorpus consists of texts from <www.yoursingapore.com>, available in Chinese, English, Indonesian, Japanese, Korean and Vietnamese. The `singaporemedicine` subcorpus comprises texts from <www.singaporemedicine.com> in Arabic, Chinese, English, Indonesian and Vietnamese.

In the initial phase we have built a corpus totaling 595,000 words (26,000 sentences) in 7 languages (Arabic, Chinese, English, Indonesian, Japanese, Korean, and Vietnamese). According to the classification in the Ethnologue (Lewis, 2009) these are from seven different language groups: (Afro-Asiatic, Sino-Tibetan, Austronesian, Indo-European, Japonic, Korean as a language isolate and Austro-Asiatic).

2.1 Crawling and Cleaning

Httrack (Roche 2007) was used for data-collection and it was completed with a single command for each website.[1]The `-p1` option of Httrack downloads only the raw HyperText Markup Language (HTML) files without the embedded media files (e.g. images, flash files, embedded videos, etc.) from the webpages.

As the markup used to construct the websites were consistent, a simple Perl script was created to extract the main body text The markup cleaning extracted the text bounded by

[1] The commands are:

a. `httrack http://www.yoursingapore.com -o +*.yoursingapore.com/content/traveller/*/*.html -p1`
b. `httrack http://www.singaporemedicine.com -o -p1`

`<p>...</p>` within the `<div class = paragraph section>...</div>` attributes. The Perl script successfully extracted the main body text from each webpage and ignored the subtexts that were headers to other pages.

Non-text characters (e.g. non-break spaces (`U+00A0`), control characters (`U+0094`)) caused errors in POS tagging and sentence alignment. A second round of cleaning removed the non-text characters before the annotation tasks. All the resulting text files were converted to and saved in the UTF-8 encoding.

2.2 Sentence Segmentation

The Arabic, Indonesian, English, Korean and Vietnamese texts use the same punctuation. We segmented them with the sent_tokenize module from the Natural Language Tool Kit (NTLK: Bird et al. 2009). The sent_tokenize program uses stop punctuations (i.e. ! ? .) to identify the end of the sentences with some exceptions for entities such as websites.

The multi-byte Chinese and Japanese sentences were separated by the same sets of ！？。 punctuation but as multi-byte characters. We used the `nltk.RegexpTokenizer (u'[^！？。]*[！？。]')` to segment the Chinese and Japanese sentences. The Japanese regex has a minor tweak from the common `nltk.RegexpTokenizer (u'[「」！？。]*[！？。]')`, as recommended by the Hagiwara's Japanese chapter of the 「入門 自然言語処理」*nyumon shizen gengo shori* "Japanese Natural Language Processing with Python" (Bird et al. 2010). The tweak was necessary to include non-sentence phrases bounded by「...」 brackets. Normally the Japanese「」brackets would have an individual sentence within the brackets. However, the text from www.yoursingapore.com used the quotes 「」 not only for sentences but also for proper names (e.g. 「マリーナ貯水池」*mari-na chosuichi* "Marina Reservoir"; 「スターバックス」*suta-bakkusu* "Starbucks") or loan phrases (e.g.「三歩一拝」*san ho ichi hai* "three step a bow" - a Chinese Buddhism term; 「ハラール」 *hara-ru* "halal"; 「カルーセル」 *karu-seru* "carousal") .

2.3 Tokenization

The tokenization (i.e. word level segmentation) tasks splits sentences up into individual "meaningful units" and these meaningful units are dependent on the philological stance of different word segmenter programs. In this paper, the term word and token will be used interchangeably to refer to the individual tokens output by the POS taggers and tokenizers.

For English and Indonesian data, whitespaces are the delimiter for the tokens. Although Vietnamese words are separated by whitespaces in the orthography, sometimes two "words" separated by whitespace are supposed to mean a single thing. For example, the Vietnamese word 'quốc tế' mean international but the individual "word" separated by the space does have its meaning ('quốc' means country and 'tế' means to run). Thus the JVnSegmenter module within JVnTextPro (Nguyen and Phan 2007) was used to tokenize the Vietnamese data.

For the Japanese and Korean word level segmentation, the segmenter is incorporated into the POS-taggers that this corpus project is using. The Arabic data was segmented using the the Stanford Arabic segmenter (Gallery and Manning 2008) according to the Arabic

TreeBank clitic segmentation and orthographic normalization standards. The Stanford Chinese word segmenter was used to segment the Chinese (Tseng et al. 2005).

In the general tourism domain the Chinese segmenter made many errors for local street names that were transliterated from English to Chinese. For example, the Stanford Chinese word segmenter wrongly tokenized 乌节路 *wujielu* “Orchard road” as 乌 节路 *wu jielu* “black joint-road”. These local terms were re-segmented with a manually crafted dictionary built using Wikipedia’s Chinese translations of English names of Singapore places and streets.

2.4 Monolingual Annotation – Part of Speech (POS) Tagging

Different programs were used to tag the individual languages with their respective POS tag sets. All the tagged output was formatted into the Corpus Work Bench (CWB) verticalized text format with eXtensible Markup Language (XML) tags to encode the start and end of a sentence (i.e. `<s>…</s>`). Table 1 presents a brief summary of the sentence segmentation and POS-tagging task for the corpus compilation.

The Arabic data was tagged using the Stanford Arabic tagger with the `arabic-accurate.tagger` model (Green and Manning 2010). The Stanford Chinese POS tagger tagged the Chinese data with the `chinese.tagger` model (Tseng et al. 2005). The Penn Arabic Treebank tagset and the Chinese Penn Treebank tagset were used by the Stanford taggers respectively.

The HunPos tagger applied the Penn Treebank II POS annotations to the English texts (Halacsy et al. 2007). The pre-trained Wall Street Journal English (`en_wsj.model`) model was used with the HunPos tagger to tag the English data.

The Indonesian data was tagged by an Indonesian POS tagger (`CRFind POS`) we re-constructed based on the state-of-the-art CRF template and Bahasa Indonesian Tagset I as described by Pisceldo et al. (2009). The tagger model was trained with the 1 million word Indonesian corpus built under the PANL Project.[2]

The Japanese data was tagged by the MeCab tagger (Kudo et al. 2004). The MeCab tagger was used with the `-Ochasen` model, which was trained by the ChaSen tagger (Matsumoto et al. 1999). Different from the other POS-tagger used in this project, the MeCab morphological analyser provided more than a layer of POS annotations; MeCab output adheres to the IPADIC 2.7.0 standards (Matsumoto and Asahara 2004).

The POSTech TAGger –Korean (POSTAG/Sejong) was used to tag the Korean text. As an agglutinative language, POSTAG/Sejong tagged the tokens at a morpheme level rather than the word level. A custom tagset with 41 tags was used by POSTAG/Sejong to suit the Korean morphemes. The POSTAG/Sejong tagger is only available on Microsoft Windows OS but we managed to run it under the WINE emulator (scripts for this are available with the corpus).

The JVnTagger (part of the JVnTextPro tool) with the `MaxEnt` model was used to annotate the Vietnamese text with the VSLP (2010) tagset.

The main problem with using the wide variety of tools was that some taggers only accept local encodings. When feeding data into the English (HunPos) and the Korean (POSTAG/Sejong) tagger, the encoding needed to be changed to the ISO-8859-1 (Latin-1) and EUC-KR (EUC Korean) respectively. This caused some problems for Korean, as the

---

[2] `http://panl10n.net/english/OutputsIndonesia2.htm`

input text contained characters that cannot be represented in the EUC-KR encoding used by POSTAG/Sejong (such as the – , é and © characters). We mapped them to - , e and (C) during the POS-tagging task for the Korean texts. We hope that more projects will produce UTF-8 versions of their morphological analyzers in the future.

Language	**Sentence Segmenter**	**Word Segmenter**	**POS-tagger (Tagger Encoding)**	**Tagset**
Arabic	NLTK sent_tokenize	Stanford Segmenter	Stanford POS tagger (UTF-8)	Penn Arabic Treebank
Chinese	NLTK RegexpTokenizer	Stanford Segmenter	Stanford POS tagger (UTF-8)	Penn Chinese Treebank
English	NLTK sent_tokenize	Whitespaces	HunPos (ISO-8859-1)	Penn Treebank II
Indonesian	NLTK sent_tokenize	Whitespaces	CRFind POS (UTF-8)	Bahasa Indoneisan Tagset I
Japanese	NLTK RegexpTokenizer	MeCab	MeCab (UTF-8)	IPAdic
Korean	NLTK sent_tokenize	POSTAG/ Sejong	POSTAG/Sejong (EUC-KR)	Sejong
Vietnamese	NLTK sent_tokenize	JVnSegmenter	JVnTagger (UTF-8)	VSLP

Table 1: Summary of Tokenization and Monolingual Annotation (POS tagging) Task

We show examples of the tagged text in Tables 2 (`yoursingapore`) and 3 (`singaporemedicine`).

Language	**Segmented, Part of Speech tagged Text**
Chinese	<s>如果_CS 您_PN 在_P 新加坡_NR 只_AD 能_VV 前往_VV 一_CD 间_M 俱乐部_NN ， _PU 祖卡_NN 酒吧_NN 必然_AD 是_VC 您_PN 的_DEG 不二_JJ 选择_NN 。_PU</s>
English	<s>If_IN you_PRP only_RB have_VBP time_NN for_IN one_CD club_NN in_IN Singapore_NN ,_, then_RB it_PRP simply_RB has_VBZ to_TO be_VB zouk_JJ ._.</s>
Indonesian	<s>Jika_nn Anda_nn hanya_rb memiliki_vbt waktu_nnc untuk_in satu_cdp klub_nnc di_in Singapura_nn ,_, pergilah_nn ke_in Zouk_nn ,_, mungkin_rb satu-satunya_jj klub_nnc malam_nn di_in Singapura_nn yang_sc bereputasi_nn internasional_jj ._.</s>
Japanese	<s>シンガポール_名詞-固有名詞-地域-国 で_助詞-格助詞-一般 一つ_名詞-一般 の_助詞-連体化 クラブ_名詞-一般 に_助詞-格助詞-一般 しか_助詞-係助詞 行く_動詞-自立 時間_名詞-副詞可能 が_助詞-格助詞-一般 なかっ_形容詞-自立 た_助動詞 と_助詞-格助詞-引用 し_動詞-自立 たら_助動詞 、_記号-読点 間違い_名詞-ナイ形容詞語幹 なく_助動詞 、_記号-読点 この_連体詞 ズーク_名詞-一般 に_助詞-格助詞-一般 行く_動詞-自立 べき_助動詞 です_助動詞 。_記号-句点 </s>
Korean	<s>싱가포르_NNP 에서_JKB 클럽_NNP 한_NNP 군데_NNB 밖에_JX 가_VV ㄹ_ETM 시간_NNG 이_JKS 없_VA 다면_EC ,_SP Zouk_SL 를_JKO 선택_NNG 하_XSV 시_EP 어요_EF ._SF</s>
Vietnamese	<s>Nếu_C bạn_N chỉ_R có_V thời_gian_N ghé_V thăm_V một_M câu_lạc_bộ_N ở_E Singapore_Np ,_, hãy_R đến_V Zouk_Np ._.</s>

Table 2: A sample of monolingual annotation from `yoursingapore`

Language	Segmented, Part of Speech tagged Text
Arabic	<s>كما_CC عمل_VBD تدفق_NN الاختصاصيين_DTNNS المشهورين_DTJJ و_CC علماء_NN الطب_DTNN البيولوجي_DTJJ الى_IN سنغافورة_NNP على_IN تعزيز_NN الخدمات_DTNNS الطبية_DTJJ التي_WP توفر_VBP ها_PRP سنغافورة_NNP اضافة_NN الى_IN الفعالية_DTNN الوظيفية_DTJJ ._PUNC</s>
Chinese	<s>知名_JJ 专家_NN 和_CC 生物_NN 医药_NN 领域_NN 的_DEG 科学家_NN 的_DEG 到来_NN ，_PU 增强_VV 了_AS 新加坡_NR 的_DEG 医疗_NN 实力_NN 和_CC 运作_NN 效能_NN 。_PU</s>
English	The_DT influx_NN of_IN renowned_JJ specialists_NNS and_CC biomedical_JJ scientists_NNS has_VBZ enhanced_VBN Singapore's_NNP medical_JJ offerings_NNS and_CC operational_JJ effectiveness_NN ._.
Indonesian	<s>Masuknya_nn tenaga_nnu spesialis_nnu dan_cc ilmuwan_nn biomedis_nn terkemuka_nn kian_nn memperkuat_vbt daya_nnu tawar_nn dan_cc efektivitas_nn operasional_jj medis_nn Singapura_nn ._.</s>
Vietnamese	<s>Singapore_Np trở_thành_V điểm_đến_N của_E rất_R nhiều_A bác_sĩ_N và_C chuyên_gia_N y_sinh_N nổi_tiếng_A và_C điều_N này_P càng_R góp_phần_V tăng_cường_V chất_lượng_N và_C hiệu_quả_N hoạt_động_N y_tế_N của_E đất_nước_N này_Np ._.</s>

Table 3: A sample of monolingual annotation from `singaporemedicine`

2.5 Monolingual Annotation – Sense Tagging

We would like to sense annotate all languages using a linked sense inventory. There are free wordnets available for Arabic, English, Japanese and Indonesian (Black et al. 2006; Fellbaum 1998; Isahara et al. 2008; Nurril Hirfana et al. 2011) and a wordnet that is free-for-research for Chinese (Xu et al 2008). Unfortunately there are currently no available wordnets for Korean and Vietnamese. Currently we have tagged Chinese and English in the `yoursingapore` domain and are in the process of tagging Indonesian and Japanese. For Chinese and Indonesian these will be the first corpora tagged with wordnet senses. While tagging, we have been providing feedback on missing senses to the upstream wordnet projects.

2.6 Cross-lingual Annotation – Sentence-level Alignment

As machine-readable dictionaries are only available for certain languages in the NTU-MC, the dictionary and length based `hunalign` tool is suitable for aligning the NTU-MC as the algorithm "remains completely meaningful even in total absence of a dictionary" (Varga et al. 2005). The alignments generated by `hunalign` are bi-directionally equivalent. The sentence-level alignment task was carried out with four different conditions:

- –dic – hunalign outputs without language pair dictionary,
- +dic – hunalign outputs with language pair dictionary,
- +human – manually aligned Gold Standard,
- +pivot – alignments generated by transitive relation using 2 +human alignments

Only sentences from the textfiles that were available in all 6 languages were sentence-aligned. Two native Chinese and Japanese speakers were enlisted to correct the +dic alignments for the English-Chinese and English-Japanese data. The English-Chinese, English-Japanese and English-Korean were generated with the `CC-CEDICT` (MDBG 2011), `JMDICT` (Breen 2004) and enhanced `engdic` (Paik and Bond 2003) respectively. By extending the idea of exploiting existing resources to building and extending valency dictionaries, we used the +human alignments to produce +pivot alignments. Using English as the pivot language, we aligned Chinese-English-Japanese.

3 Corpus Evaluation

The corpus evaluation is based on the data availability, corpus outputs and its monolingual and cross-lingual annotations. The monolingual annotations were evaluated extrinsically by measuring Inter-annotator Agreement (IAA) between the POS-taggers and human annotators. Because we were using so many different parsers for so many different languages we could not tag a gold standard for each language. The quality of the parallel text alignments was intrinsically evaluated by computing the F-score of the `hunalign` outputs against manually aligned data.

3.1 Corpus Availability

For a corpus to be a valuable resource, it must be both useful and accessible (Ishida 2006). The owners of the source data (Singapore Tourism Board) have allowed the redistribution of this data, licensed by the Creative Commons (CC) Attribution 3.0 Unported License. Users of the corpus are able to share (i.e. copy, distribute and transmit) and remix (i.e. to adapt) the corpus under the condition of attributing the work to the NTU-MC project. The data is available from the project website: http://linguistics.hss.ntu.edu.sg/ResearchinLMS/Pages/NTUMultilingualCorpus.aspx.

3.2 Corpus Size

The NTU-MC project compiled a foundation text of 595,000 words (26,000 sentences) for the NTU-MC in 7 languages from 7 language family trees. The breakdown of the monolingual annotation is as followed (the number. of tokens excludes punctuations and symbols, the number of concepts includes open class words not found in wordnet):

	Language (language code)		**Language Family**	**#Texts**	**#Sents**	**#Tokens**	**#Concepts**
yoursingapore subcorpus	Chinese	(cmn)	Sino-Tibetan	280	2,365	52,047	41,186
	English	(eng)	Indo-European	398	3,255	76,339	43,990
	Indonesian	(ind)	Austronesian	270	2,185	50,315	38,102
	Japanese	(jpn)	Japonic	267	2,648	72,797	43,227
	Korean	(kor)	Language Isolate	266	2,407	67,341	
	Vietnamese	(vie)	Austro-Asiatic	269	2,236	56,535	
singapore-medicin	Arabic	(msa)	Afro-Asiatic	73	1,909	46,222	
	Chinese	(cmn)	Sino-Tibetan	70	1,760	38,994	
	English	(eng)	Indo-European	118	3,801	71,598	
	Indonesian	(ind)	Austronesian	71	1,789	26,687	

	Vietnamese (vie)	Austro-Asiatic	72	1,838	35,628	
	Total:	7 Families	2,154	26,193	594,503	

Table 4: Monolingual Annotation Outputs

Another way of looking at it is there are roughly 3,900 sentences of aligned text, with 2,200 having six languages and 1,700 having six languages. Both sets have Chinese, English, Indonesian and Vietnamese.

Presently, cross-lingual annotations are only available for the `yoursingapore` subcorpus. The main alignment task for NTU-MC focused on the English-Asian Languages alignments due to the amount of lexical resources available for English bitext. The corpus produced 2 Gold Standard (+human) alignments, 3 +dic alignments, 1 +pivot alignment and 11 -dic alignments generated with the `null.dic` option on hunalign.

	eng	cmn	ind	jpn	kor	vie
eng						
cmn	+human / +dic					
ind	−dic	−dic				
jpn	+human / +dic	+pivot	−dic			
kor	+dic	−dic	−dic	−dic		
vie	−dic	−dic	−dic	−dic	−dic	

Table 5: Cross-lingual Annotation Outputs (`yoursingapore` subcorpus)

3.3 Monolingual Annotation Evaluation

One text from each subcorpus was selected at random for human annotators to verify the POS-taggers' accuracy; the `fish-head-curry.txt` from `yoursingapore` and the `leadingmedhub1.txt` from `singaporemedicine` subcorpus. The human annotators were assigned to verify the POS tags and mis-segmented tokens. The accuracy of the human annotation might be primed by what the POS tagger had tagged. Therefore the human verifications were not treated as the "gold standard" but an inter-annotation agreement (IAA) score that was derived from the annotators' identification of the mis-segmented and mis-tagged tokens.[3]

For the Japanese POS evaluation, there was no human annotator available. Thus a different POS tagger, ChaSen morphemic analyzer, was used to calculate IAA. Both programs uses the ipadic POS, but the noticeable difference is that ChaSen is more conservative when tagging unknown words: ChaSen applied the 未知 *michigo* "unknown word" tag to tokens for unseen words whereas MeCab forces the closest fit POS to the unknown tokens. The 12 instances of 未知語 tags in `fish-head-curry.txt` were not included in the IAA calculation.

The low IAA score for the Indonesian POS tags was partially because the human annotator penalized the POS tagger for tagging most proper nouns (NNP) with the common noun (NN) tag; 19 out of the 104 mis-tags were of this nature. Disregarding the NNP penalty, the IAA would have been 78.26%.

[3] This excludes punctuation and both the number of mis-segments and mis-tagged tokens.

Language	Sentence Order	#Tokens	#Sentences	#Mis-segments	#Mis-tagged	IAA	Reported accuracy
msa[4]	SVO	217	8	1	8	95.85%	95.58% (Green and Manning, 2010)
cmn	SVO	401	16	20	23	92.29%	93.65% (Tseng et al, 2005)
eng	SVO	410	14	-	23	94.39%	96.58% (Halacsy et al, 2009)
ind	SVO	391	15	-	104	73.40%	83.72% (Pisceldo et al, 2009)
jpn[5]	SOV	293	14	3	8	96.25%	97.66 % (Kudo et al, 2004)
kor[5]	SOV	374	14	44	27	81.02%	90.7% (Lee et al, 2002)
vie	SVO	420	14	17	23	90.48%	93.32% (Nguyen et al, 2010)

Table 6: Summary of Segmentation and POS Annotation Task

The IAA reported in table 4 serves as a gauge, an error bar, of the reported accuracy reported by the individual taggers. The IAA score accounts for counts from both `fish-head-curry.txt` and the `leadingmedhub1.txt` whenever possible. The IAA is measured as such:

non-matches = no. of mis-segment + no. of mis-tagged (1)
matches = no. of tokens – non-matches (2)
IAA = matches / (matches + non-matches) * 100% (3)

3.4 Cross-lingual Annotation Evaluation

A subset of 9 text files was selected to evaluate the quality of the `hunalign` outputs for language pairs with English sentences. The evaluation metrics adheres to standards set by the ARCADE II project (Chiao et al. 2006); the recall, precision and F-score is computed on the `hunalign` output of word segmented sentences. F-scores were computed using sentence and character granularity (with and without spaces).

From Figure 1, the alignment task on Japanese, Korean and Chinese is a much more difficult task than aligning Indonesian or Vietnamese data; even with the dictionaries' input, alignments for non-Latin character-based languages are poorer in alignments. Possibly, it is the difference in sentence order (refer table 4) that affected the lexicon quality of the Japanese-English and Korean-English alignments. Nevertheless, +human alignments were manually crafted for English-Japanese and English-Chinese sentences and the English-Korean alignment is reasonably good in terms of character granularity.

[4] Only from `leadingmedhub1.txt`
[5] Only from `fish-head-curry.txt`

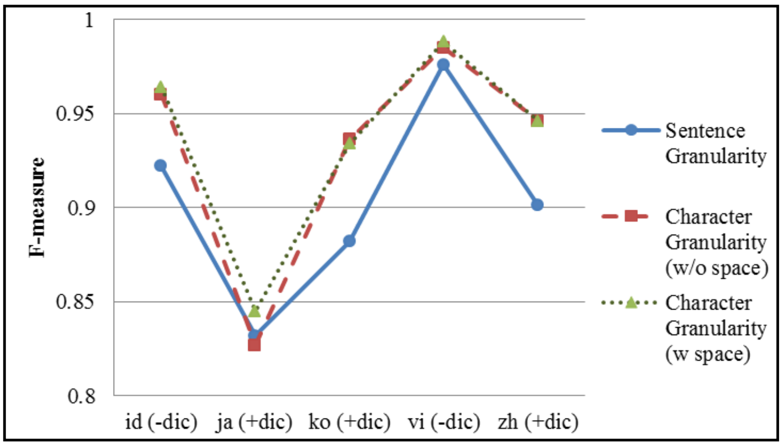

Figure 1: F-measure of hunalign on English-Asian Language alignments

The primary advantage of pivoting alignments to generate other language-pairs alignments is the simplicity to leverage on Gold Standard alignments to produce alignments where the bilinguals of the language pairs are scarce. Similar to the idea of increasing the number of language pairs quadratically by sourcing parallel sources with more languages (Eisele and Chen 2010), +pivot alignments can produce +human like alignments quadratically with each +human alignments. Although it is possible to create more alignments through other pivoting permutations, generating pivoted alignments from crude -dic alignments will be perpetuating the original mis-alignments that `hunalign` had produced. Thus only the pivoted Gold Standard alignments was worth the effort as it can be able to produce word-level alignments of similar quality to the +human alignments.

4 Discussion and Work in Progress

A comparable corpus to the NTU-MC is OPUS which taps open source parallel text (Tiedemann 2009). The OPUS is representative of a global open source enthusiast's community, while the NTU-MC targets data from a specific cosmopolitan society. The OPUS covers a wider range of domains with large sub-corpora and it provides automated monolingual (POS tags and syntactic parses) and cross-lingual (sentence and word level alignments) annotations; whereas the NTU-MC is a corpus of a smaller size but more diverse in Asian language data. Over time we intend to annotate more phenomena and create a multilingual Gold Standard annotation beneficial for a variety of NLP tasks.

The NTU-MC is the focus of an ongoing effort to add content, layers of annotation and usability as it continues to make multilingual resources machine readable for NLP tasks. Current work on the NTU-MC involves increasing both the amount of data, and the richness of the monolingual and cross-lingual annotations.

We were also granted the permission to use parallel text (English, Malay, Chinese and Tamil) distributed by National Environment Agency of Singapore (NEA) and Sembawang Town Council (SBTC). However these texts are embedded in image formats or flattened portable document format thus text extraction is dependent on manual input or Optical Character Recognition (OCR) technologies of different languages. Several attempts to use open/free OCR software resulted in noisy text outputs that requires much cleaning. These texts from NEA and SBTC will be used in future extension of the corpus when resources allow for manual data entry (e.g. mechanical turks) or proprietary OCR software for Asian languages that performs reasonably well. Also, we are constantly requesting for parallel public informational text from other governmental authorities.

Although we have exploited prior knowledge put into the design of the POS tag sets and token segmentations using different (ad-hoc) tools, the philological perspective on segmentations and POS varies within each individual language and across languages. To fill these philological and cross-lingual gaps in the monolingual annotations, we are working to provide syntactic annotation with the Deep Linguistic Processing with HPSG Initiative (DELPH-IN)[6] and semantic annotation with the Global WordNet Association (GWA).[7] From the parses of the individual languages, the multi-layered annotation will allow extraction of the syntactic annotations (e.g. POS from HPSG word classes, word boundary from HPSG lexicon) and semantic annotations (e.g. semantic constraints from HPSG lexicon and its corresponding word senses mapped to WordNet). Wordnet sense annotation of the Indonesian and Japanese data from the `yoursingapore` subcorpus is ongoing.

For cross-lingual annotation, sentence-level, word-level and concept-level alignment will be carried out as resources permit. These word alignments from the hitherto under-represented language pairs should provide rich data for language technologies like MT and IR.

The NTU-MC is being used as a teaching tool, both in courses on corpus linguistics and semantics and as material for student projects. In the semantics class, students annotated short tourism pages (three students to a page) then looked at their inter-annotator agreement and reported on words where they had disagreed as to the correct sense as well as on words missing from the sense inventory (Princeton Wordnet). Students said that they found the concrete task interesting and that it really made the issues involved in defining word meanings clear. A similar task was done on the Chinese portion for a class in Chinese lexicography. When the corpus has been checked once more we intend to submit it as a sense tagged corpus multi-text to the Natural Language Tool Kit.

5 Conclusion

This project has produced a text collection, the NTU Multilingual Corpus, small in size but rich in language diversity. The NTU-MC contains a layer of monolingual annotation (POS tags and some sense tags) as well as a layer of cross-lingual annotation (sentence-level alignments) valuable for cross-lingual NLP tasks. The texts and annotation are released under an open license (CC by). In a cosmopolitan city like Singapore, there is a wealth of parallel text. This project urges future research to continue to draw diverse data through readily available yet untapped resources for corpus compilation. By progressively extending the NTU-MC with a larger dataset and multiple layers of annotation, it expands the scope of the usage and becomes a better corpus for general or computational linguistics researches. By building corpora of more diverse cross-lingual nature, it provides information on the unique sociolinguistic situation in linguistically diverse societies (e.g. translatability researches, language choice and language domain researches); also it pushes

[6] http://www.delph-in.net/
[7] http://www.globalwordnet.org/

the state-of-the-art NLP techniques through more robust cross-lingual training (Matsumoto et al. 1993).

6 Acknowledgement

The authors of this paper thank Ms Joan Lee, New Media Manager of the Singapore Tourist Board, for granting permission to use and redistribute text from their multilingual websites (www.yoursingapore.com.sg, www.singaporemedicine.com, http://app.singaporeedu.gov.sg/asp/index.asp). The authors also thank Ms Dorothy Cheung, Public Relations Manager of Sembawang Town Council (SBTC) and Mr Edrick Chua, Assistant Director of Corporate Communications from National Environment Agency (NEA) for their permission and aid in providing access to their data. Though the data from SBTC and NEA is not used for the current phase of NTU-MC compilation, we hope to use it for the future extension of the corpus.

This research was partially funded by a joint JSPS/NTU grant on Revealing Meaning through Multiple Languages and the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2009-5259-5.

7 References

Bird, S., Klein, E., Loper, E., 萩原正人 (Hagiwara, M.), 中山敬広 (Nakayama, T.) and 水野貴明 (Mizuno, T.) (translation), 2010, 入門 自然言語処理 (Introduction to Natural Language Processing), O’Reilly, Japan (translation, with one extra chapter, of Bird *et al.* 2009).

Bird, S., Ewan, K., and Loper, E, 2009, Natural Language Processing with Python, O'Reilly Media.

Black W., Elkateb S., Rodriguez H., Alkhalifa M., Vossen P., Pease A., Bertran M., Fellbaum C, 2006, The Arabic WordNet Project, *Proceedings of LREC 2006.*

Breen, J.W. 2004, JMDict: a Japanese-multilingual dictionary, In *Coling 2004 Workshop on Multilingual Linguistic Resources*, Geneva, pp. 71–78.

Chiao, Y.C., Kraif, O., Laurent, D., Nguyen, T.M.H., Semmar, N., Stuck, F., Veronis, J. and Zaghouani,W., 2006, Evaluation of multilingual text alignment systems: the ARCADE II project., *Proceedings of the LREC 2006 Conference*.

Eisele, A. and Chen, Y., 2010, MultiUN: A multilingual corpus from United Nation documents, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10)*, Valletta, Malta.

Erjavec, T, 2004, MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora, *Fourth International Conference on Language Resources and Evaluation, LREC'04*, ELRA, Paris. pp. 1535-1538.

Christiane Fellbaum. (ed.), 1998, *WordNet: An Electronic Lexical Database*, MIT Press.

Galley, M. and Manning, C.D., 2008, A Simple and Effective Hierarchical Phrase Reordering Model, In *Proceedings of Association for Computational Linguistics*, Ohio, USA.

Green, S. and Manning, C.D., 2010. Better Arabic Parsing: Baselines, Evaluations, and Analysis, In *Proceedings of International Conference on Computational Linguistics (COLING)*, Beijing, China.

Halácsy, P., Kornai, A. and Oravecz, C., 2007, HunPos - an open source trigram tagger 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.209--212.

Hwa, R., Resnik, R., Weinberg, A., Cabezas, C., and Kolak,C., 2005, Bootstrapping parsers via syntactic projection across parallel texts, *Natural Language Engineering*, 11(3), pp. 311–325.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M. and Kanzaki, K., 2008, Development of Japanese WordNet, In *LREC-2008*, Marrakech.

Ishida, T., 2006, Language Grid: An Infrastructure for Intercultural Collaboration, In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp.96-100, keynote.

Kudo T., Yamamoto, K., and Matsumoto, Y., 2004, Applying conditional random fields to Japanese morphological analysis, In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pp.230–237.

Lewis, P.M. 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: *http://www.ethnologue.com/*.

Matsumoto, Y., Ishimoto, H. and Utsuro, T., 1993, Structural matching of parallel texts. In *31st Annual Meeting of the Association for Computational Linguistics: ACL-93*, pp.23–30.

Matsumoto, Y., Takaoka, K. and Asahara, M., 1999, ChaSen Morphological Analyzer version 2.4.0 User's Manual. NAIST Technical Report, Nara Institute of Science and Technology Technical Report 9009, Retrieved on 07 Jan 2011 from *http://sourceforge.jp/projects/chasen-legacy/docs/chasen-2.4.0-manual-en.pdf/en/1/ chasen-2.4.0-manual-en.pdf.pdf*

McCallum, B., 2011, Translation Technology at the United Nations, *MultiLingual Computing & Technology*, 15(2), p. 62.

MDGB, 2011, CC-CEDICT [Machine-Readable Dictionary]. Netherlands : MDGB, Retreived May 03, 2011 from *http://www.mdbg.net/chindict/chindict.php?page=cedict*.

Mitamura, T., Lin, F., Shima, H., Wang, M., Ko, J., Betteridge, J., Bilotti, M., Schlaikjer, A., and Nyberg, E., 2007, JAVELIN III: Cross-Lingual Question Answering from Japanese and Chinese Documents, *Proceedings of NTCIR-6 Workshop Meeting*, Tokyo, Japan.

Nguyen, C.T. and Phan, X.H. 2007. JVnSegmenter: A Java-based Vietnamese Word Segmentation Tool. Retrieved on 30 Jan 2011 from http://jvnsegmenter.sourceforge.net/

Nurril Hirfana, M.N., Sapuan, S., and Bond, F., 2011, Creating the open Wordnet Bahasa In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)* pages 258–267. Singapore.

Paik, K. and Bond, F., 2003, Enhancing an English/Korean Dictionary, In *Papillon 2003 Workshop on Multilingual Lexical Databases*, Sapporo, Japan.

Pisceldo, F., Manurung, R., and Mirna, A., 2009, Probabilistic Part-of-Speech Tagging for bahasa Indonesia, In *Third International MALINDO Workshop*, colocated event *ACL-IJCNLP 2009*, Singapore.

Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D., 2006, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy.

Resnik, P., 2004, Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation, In Alexander Gelbukh (ed.), Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing, Springer, 2004, pp. 283-299.

Roche, X., 2007, Httrack Website Copier - Offline Browser. [Computer Software]. Retrieved Jan 30, 2011, Available from *http://www.httrack.com/*.

Sarrafzadeh, B., Yakovets, N., Cercone, N., & An, A., 2011, Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources, (Technical Report CSE-2011-01). Ontario: Department of Computer Science and Engineering.

Saravanan, K., Udupa, R., and Kumaran, A., 2010, Crosslingual Information Retrieval System Enhanced with Transliteration Generation and Mining, In *Forum for Information Retrieval Evaluation (FIRE-2010) Workshop*, Kolkata, India.

Snyder, B. and Barzilay, R., 2008, Cross-lingual propagation for morphological analysis, In *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*, pp. 848–854.

Tiedemann, J., 2009, News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces, In *Proceedings of RANLP'09*, pp.237–248, Borovets, Bulgaria.

Tseng, H., Jurafsky, D. and Manning, C., 2005, Morphological features help POS tagging of unknown words across language varieties, In *Proceedings of the Fourth SIGHAN Workshop*, Jeju Island, Korea.

Varga, D., Nemeth, L., Halacsy, P., Kornai, A., Tron, V. and Nagy, V., 2005, Parallel corpora for medium density languages, In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*. pp.590–596. Borovets. Bulgaria.

VLSP Project, 2010, VLSP Project – Vietnamese Language Processing. Retrieved on 02 Jan 2010 from http://vlsp.vietlp.org:8080/demo/?page=about .

Xiao, Z., McEnery, A., Baker, P. and Hardie, A., 2004, Developing Asian language corpora: standards and practice. *Proceedings of the 4th Workshop on Asian Language Resources*, Sanya, Hainan Island. pp. 1-8.

Renjie Xu, Zhiqiang Gao, Yuzhong Qu, and Zhisheng Huang, 2008, An Integrated Approach for Automatic Construction of Bilingual Chinese-English WordNet In *Proceedings of the 3rd Asian Semantic Web Conference (*ASWC 2008*)*, Bangkok, Thailand

Zhang, Y., Uchimoto, K., Ma, Q. and Isahara, H., 2005, Building an Annotated Japanese-Chinese Parallel Corpus – A Part of NICT Multilingual Corpora, In *Second International Joint Conference on Natural Language Processing*, Jeju Island, Republic of Korea. pp 85-90.