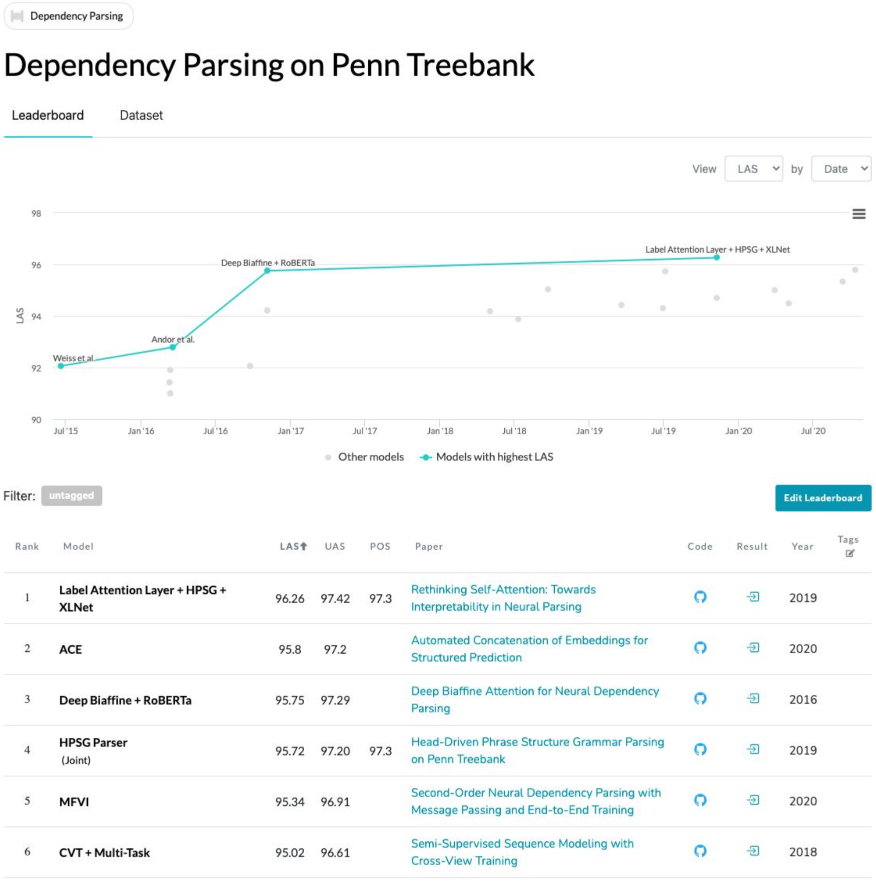
Project SEACoreNLP

Building Natural Language Processing Resources for ASEAN Languages as a Region

AI Singapore, Singapore

Leong Wei Qi and Dr. William Tjhi

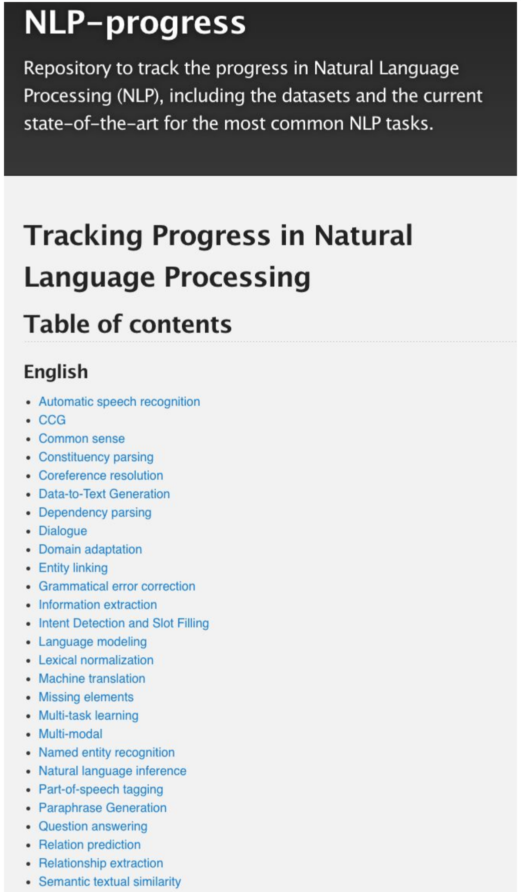
Background



Dependency Parsing on Penn Treebank

Rank	Model	LAS	UAS	POS	Paper	Code	Result	Year	Tags
1	Label Attention Layer + HPSG + XLNet	96.26	97.42	97.3	Rethinking Self-Attention: Towards Interpretability in Neural Parsing			2019	
2	ACE	95.8	97.2		Automated Concatenation of Embeddings for Structured Prediction			2020	
3	Deep Biaffine + RoBERTa	95.75	97.29		Deep Biaffine Attention for Neural Dependency Parsing			2016	
4	HPSG Parser (Joint)	95.72	97.20	97.3	Head-Driven Phrase Structure Grammar Parsing on Penn Treebank			2019	
5	MFVI	95.34	96.91		Second-Order Neural Dependency Parsing with Message Passing and End-to-End Training			2020	
6	CVT + Multi-Task	95.02	96.61		Semi-Supervised Sequence Modeling with Cross-View Training			2018	

https://paperswithcode.com/sota/dependency-parsing-on-penn-treebank



NLP-progress

Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

Tracking Progress in Natural Language Processing

Table of contents

English

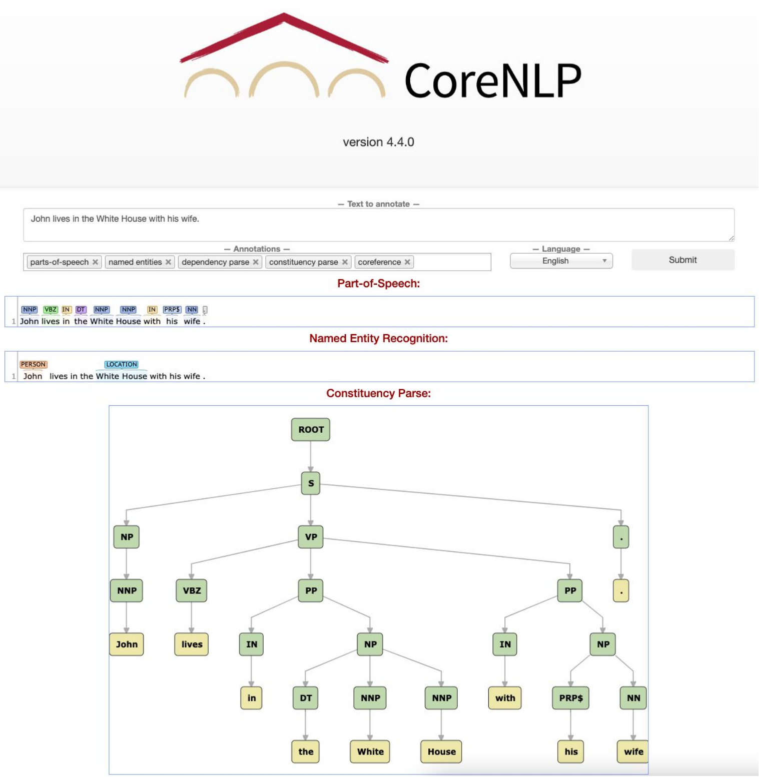
http://nlpprogress.com/

Targets

1. To build open high-quality benchmark datasets in official ASEAN languages for a core set of NLP tasks for training, evaluation and probing

2. To catalyze the use of NLP capabilities in these languages in the industry by showcasing them on a demo website and making them easy to use with a Python package

3. To establish a regional coalition for knowledge and resource sharing for NLP in ASEAN languages



CoreNLP

version 4.4.0

— Text to annotate —

John lives in the White House with his wife.

— Annotations —

parts-of-speech ✕ named entities ✕ dependency parse ✕ constituency parse ✕ coreference ✕

— Language —

English

Part-of-Speech:

```
  NNP  VBZ IN DT  NNP   NNP  IN  PRP$ NN  .
1 John lives in the White House with his wife .
```

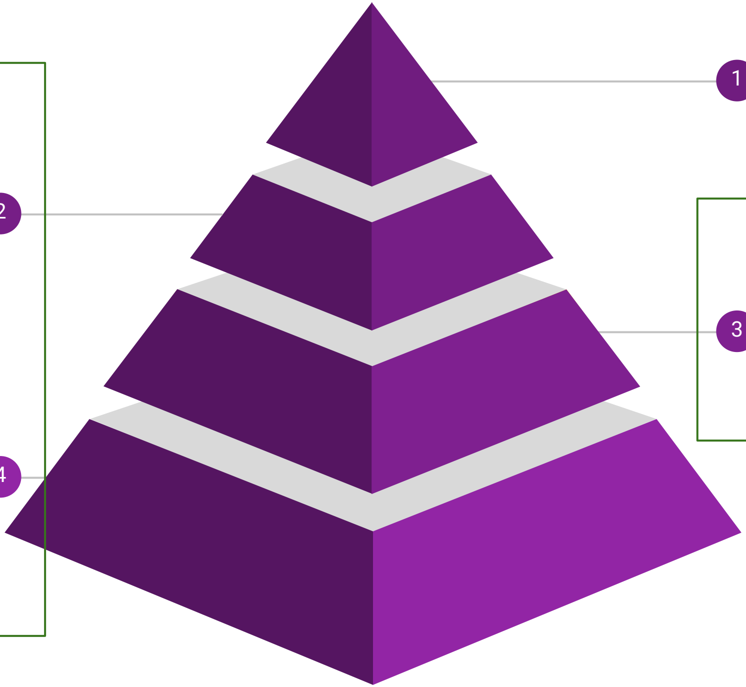
Named Entity Recognition:

```
  PERSON          LOCATION
1 John  lives in the White House with his wife .
```

Constituency Parse:

https://stanfordnlp.github.io/CoreNLP/demo.html

Proposed Solution - Benchmark Datasets/ Tools for Core NLP



1 **Natural Language Understanding/Generation**

Machine Translation, Relation Extraction, Text Summarization/Paraphrasing, Question Answering

2 **Discourse/ Semantic Analysis**

Coreference Resolution, Semantic Role Labelling, Named Entity Recognition

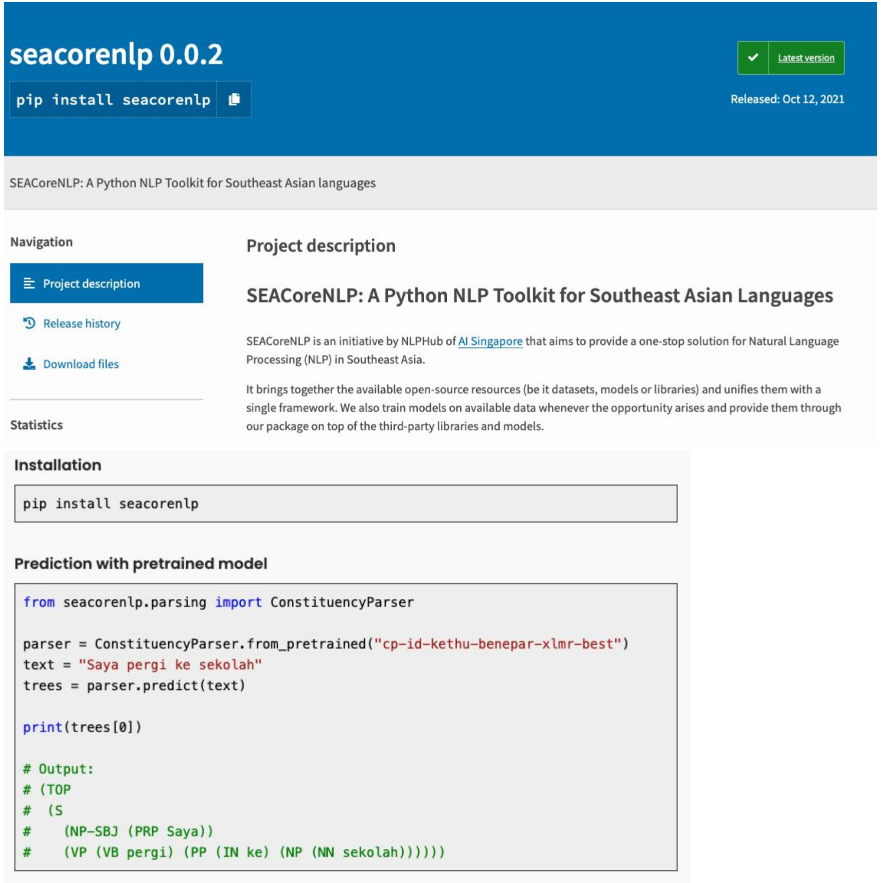
3 **Syntactic Analysis (Parsing)**

Constituency Parsing, Dependency Parsing

4 **Morphological Analysis (Segmentation, Tagging)**

Tokenization, Sentence Segmentation, Stemming, Lemmatization, POS Tagging

Proposed Solution - Python Package & Documentation



seacorenlp 0.0.2

```
pip install seacorenlp
```

Released: Oct 12, 2021

SEACoreNLP: A Python NLP Toolkit for Southeast Asian languages

Project description

SEACoreNLP: A Python NLP Toolkit for Southeast Asian Languages

SEACoreNLP is an initiative by NLPHub of AI Singapore that aims to provide a one-stop solution for Natural Language Processing (NLP) in Southeast Asia.

It brings together the available open-source resources (be it datasets, models or libraries) and unifies them with a single framework. We also train models on available data whenever the opportunity arises and provide them through our package on top of the third-party libraries and models.

Installation

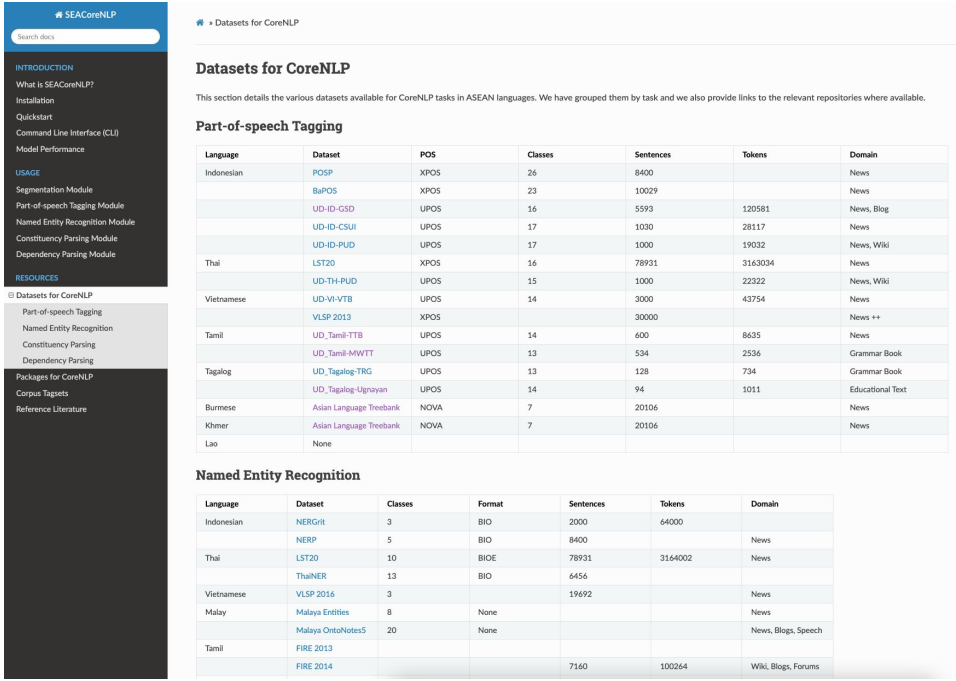
```
pip install seacorenlp
```

Prediction with pretrained model

```
from seacorenlp.parsing import ConstituencyParser

parser = ConstituencyParser.from_pretrained("cp-id-kethu-benepar-xlmr-best")
text = "Saya pergi ke sekolah"
trees = parser.predict(text)

print(trees[0])

# Output:
# (TOP
#  (S
#    (NP-SBJ (PRP Saya))
#    (VP (VB pergi) (PP (IN ke) (NP (NN sekolah))))))
```

Documentation

- Information on datasets, models, tools
- Annotation guidelines, Tagset information
- API Documentation for Python package



Datasets for CoreNLP

This section details the various datasets available for CoreNLP tasks in ASEAN languages. We have grouped them by task and we also provide links to the relevant repositories where available.

Part-of-speech Tagging

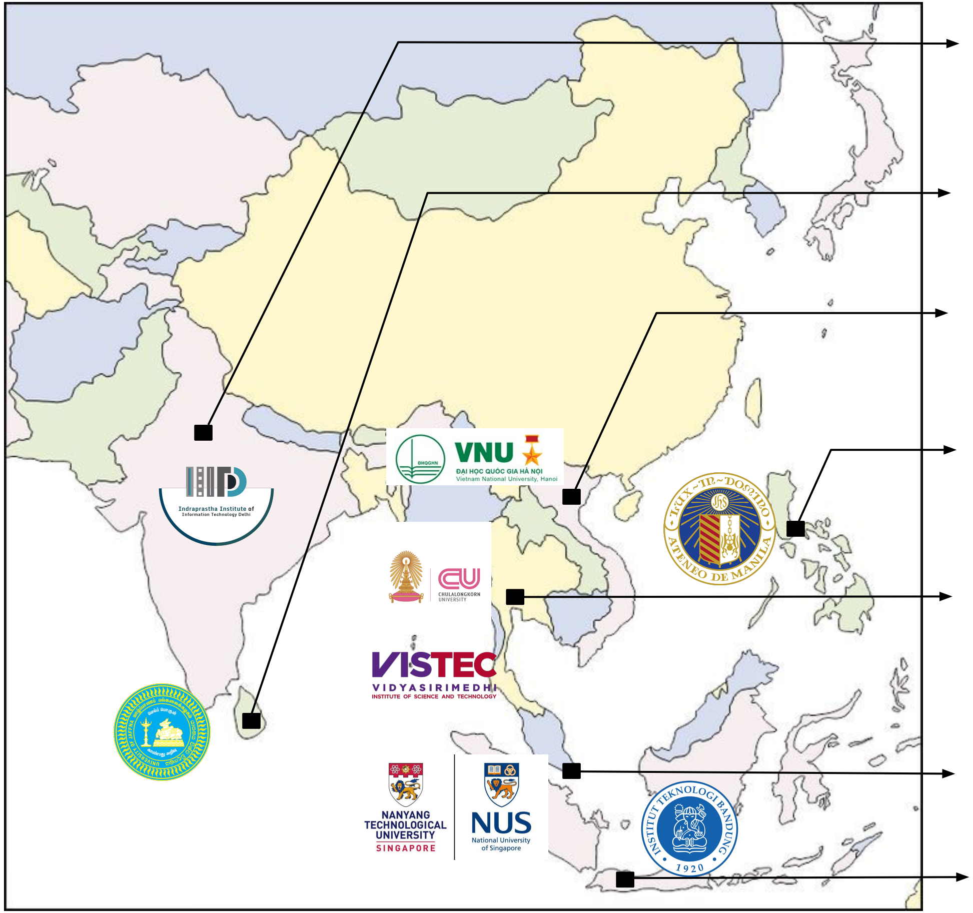
Language	Dataset	POS	Classes	Sentences	Tokens	Domain
Indonesian	POSP	XPOS	26	8400		News
	BaPOS	XPOS	23	10029		News
	UD-ID-GSD	UPOS	16	5593	120581	News, Blog
	UD-ID-CSUI	UPOS	17	1030	28117	News
	UD-ID-PUD	UPOS	17	1000	19032	News, Wiki
Thai	LST20	XPOS	16	78931	3163034	News
	UD-TH-PUD	UPOS	15	1000	22322	News, Wiki
Vietnamese	UD-VI-VTB	UPOS	14	3000	43754	News
	VLSP 2013	XPOS		30000		News ++
Tamil	UD_Tamil-TTB	UPOS	14	600	8635	News
	UD_Tamil-MWTT	UPOS	13	534	2536	Grammar Book
Tagalog	UD_Tagalog-TRG	UPOS	13	128	734	Grammar Book
	UD_Tagalog-Ugnayan	UPOS	14	94	1011	Educational Text
Burmese	Asian Language Treebank	NOVA	7	20106		News
Khmer	Asian Language Treebank	NOVA	7	20106		News
Lao	None					

Named Entity Recognition

Language	Dataset	Classes	Format	Sentences	Tokens	Domain
Indonesian	NERGrit	3	BIO	2000	64000	
	NERP	5	BIO	8400		News
Thai	LST20	10	BIOE	78931	3164002	News
	ThaiNER	13	BIO	6456		
Vietnamese	VLSP 2016	3		19692		News
Malay	Malaya Entities	8	None			News
	Malaya OntoNotes5	20	None			News, Blogs, Speech
Tamil	FIRE 2013					
	FIRE 2014			7160	100264	Wiki, Blogs, Forums

https://seacorenlp.aisingapore.net/docs/

Proposed Solution - Regional NLP Coalition


Indraprastha Institute of Information Technology Delhi
VNU
ĐẠI HỌC QUỐC GIA HÀ NỘI
Vietnam National University, Hanoi
ATENEO DE MANILA
CU
CHULALONGKORN UNIVERSITY
VISTEC
VIDYASIRIMEDHI
INSTITUTE OF SCIENCE AND TECHNOLOGY
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE
NUS
National University of Singapore
INSTITUT TEKNOLOGI BANDUNG
1920

India

- Indraprastha Institute of Information Technology (IIIT-Delhi)

Sri Lanka

- University of Jaffna

Vietnam

- Vietnam National University, Hanoi (VNU)

The Philippines

- Ateneo de Manila University

Thailand

- Chulalongkorn University (CU)
- Vidyasirimedhi Institute of Science and Technology (VISTEC)

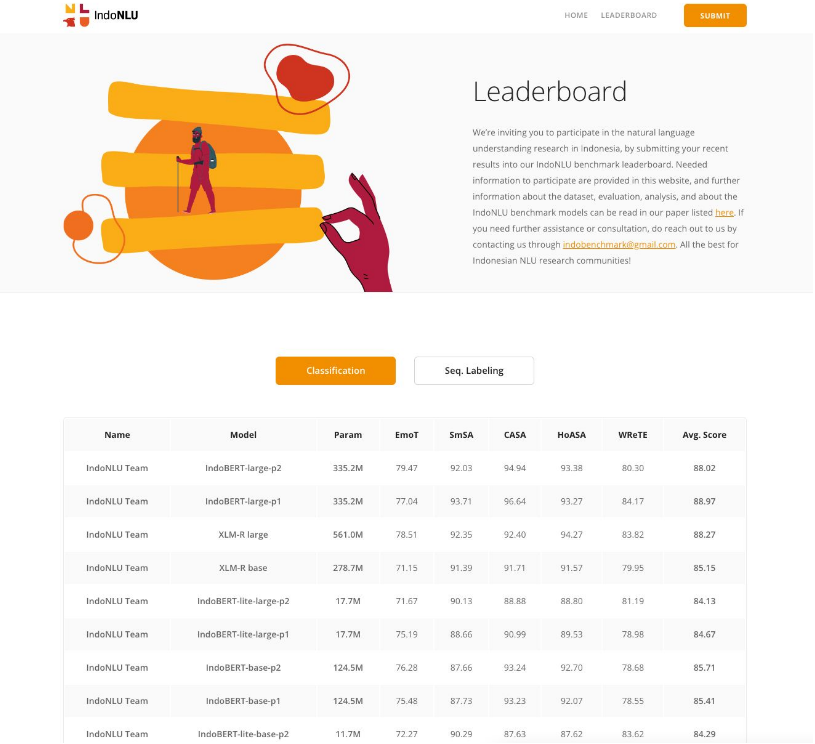
Singapore

- National University of Singapore (NUS)
- Nanyang Technological University (NTU)

Indonesia

- Institut Teknologi Bandung (ITB)

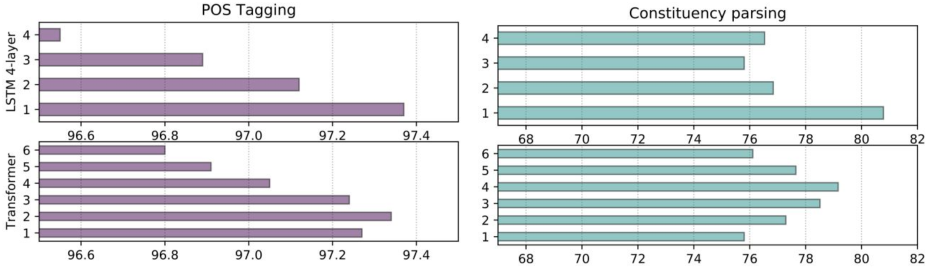
Impact - Datasets



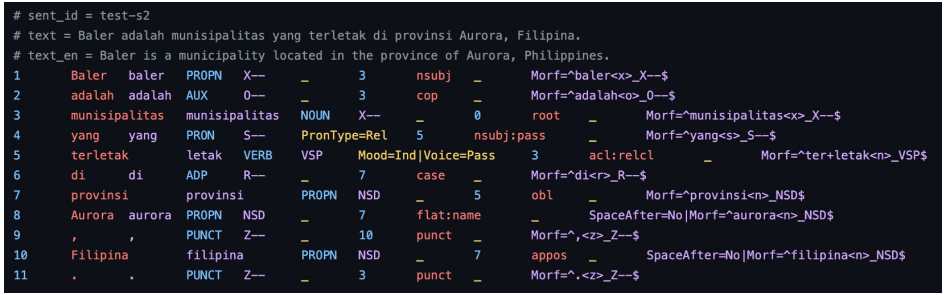
Leaderboard

We're inviting you to participate in the natural language understanding research in Indonesia, by submitting your recent results into our IndoNLU benchmark leaderboard. Needed information to participate are provided in this website, and further information about the dataset, evaluation, analysis, and about the IndoNLU benchmark models can be read in our paper listed here. If you need further assistance or consultation, do reach out to us by contacting us through indobenchmark@gmail.com. All the best for Indonesian NLU research communities!

Name	Model	Param	EmoT	SmSA	CASA	HoASA	WReTE	Avg. Score
IndoNLU Team	IndoBERT-large-p2	335.2M	79.47	92.03	94.94	93.38	80.30	88.02
IndoNLU Team	IndoBERT-large-p1	335.2M	77.04	93.71	96.64	93.27	84.17	88.97
IndoNLU Team	XLM-R large	561.0M	78.51	92.35	92.40	94.27	83.82	88.27
IndoNLU Team	XLM-R base	278.7M	71.15	91.39	91.71	91.57	79.95	85.15
IndoNLU Team	IndoBERT-lite-large-p2	17.7M	71.67	90.13	88.88	88.80	81.19	84.13
IndoNLU Team	IndoBERT-lite-large-p1	17.7M	75.19	88.66	90.99	89.53	78.98	84.67
IndoNLU Team	IndoBERT-base-p2	124.5M	76.28	87.66	93.24	92.70	78.68	85.71
IndoNLU Team	IndoBERT-base-p1	124.5M	75.48	87.73	93.23	92.07	78.55	85.41
IndoNLU Team	IndoBERT-lite-base-p2	11.7M	72.27	90.29	87.63	87.62	83.62	84.29



2 Allow for probing of language models[2]



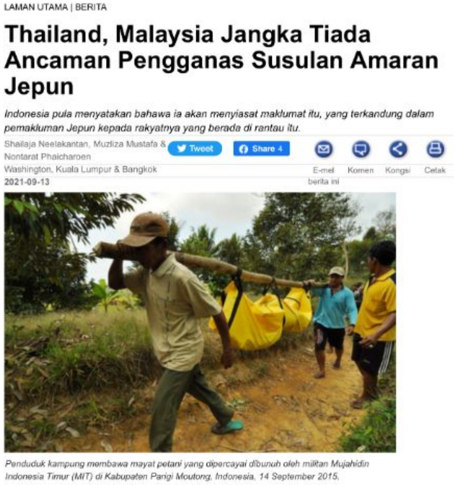
```
# sent_id = test-s2
# text = Baler adalah munisipalitas yang terletak di provinsi Aurora, Filipina.
# text_en = Baler is a municipality located in the province of Aurora, Philippines.
1       Baler   baler   PROPN   X--     _       3       nsubj   _       Morf=^baler<x>_X--$
2       adalah  adalah  AUX     O--     _       3       cop     _       Morf=^adalah<o>_O--$
3       munisipalitas   munisipalitas   NOUN    X--     _       0       root    _       Morf=^munisipalitas<x>_X--$
4       yang    yang    PRON    S--     PronType=Rel    5       nsubj:pass      _       Morf=^yang<s>_S--$
5       terletak        letak   VERB    VSP     Mood=Ind|Voice=Pass     3       acl:relcl       _       Morf=^ter+letak<n>_VSP$
6       di      di      ADP     R--     _       7       case    _       Morf=^di<r>_R--$
7       provinsi        provinsi        PROPN   NSD     _       5       obl     _       Morf=^provinsi<n>_NSD$
8       Aurora  aurora  PROPN   NSD     _       7       flat:name       _       SpaceAfter=No|Morf=^aurora<n>_NSD$
9       ,       ,       PUNCT   Z--     _       10      punct   _       Morf=^,<z>_Z--$
10      Filipina        filipina        PROPN   NSD     _       7       appos   _       SpaceAfter=No|Morf=^filipina<n>_NSD$
11      .       .       PUNCT   Z--     _       3       punct   _       Morf=^.<z>_Z--$
```

3 Promote linguistic research[3]

1 Benchmark Leaderboard → Catalyze research and development[1]

1. https://www.indobenchmark.com/
2. https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf
3. https://github.com/UniversalDependencies/UD_Indonesian-GSD/blob/master/id_gsd-ud-test.conllu

Impact - Trained Models



LAMAN UTAMA | BERITA

Thailand, Malaysia Jangka Tiada Ancaman Pengganas Susulan Amaran Jepun

Indonesia pula menyatakan bahawa ia akan menyiasat maklumat itu, yang terkandung dalam pemakluman Jepun kepada rakyatnya yang berada di rantau itu.

Shailaja Neelakantan, Muzliza Mustafa & Nontarat Phaicharoen
Washington, Kuala Lumpur & Bangkok
2021-09-13

Penduduk kampung membawa mayat petani yang dipercayai dibunuh oleh militan Mujahidin Indonesia Timur (MIT) di Kabupaten Parigi Moutong, Indonesia, 14 September 2015.

Summary (extractive)

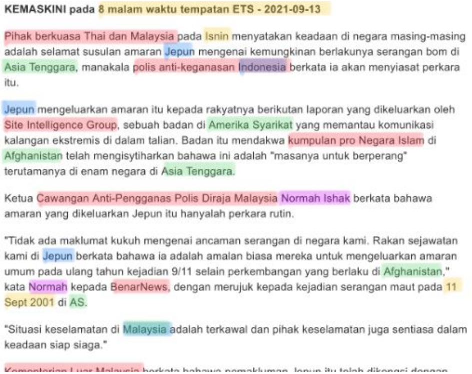
Pihak berkuasa Thai dan Malaysia pada Isnin menyatakan keadaan di negara masing-masing adalah selamat susulan amaran Jepun mengenai kemungkinan berlakunya serangan bom di Asia Tenggara, manakala polis anti-keganasan Indonesia berkata ia akan menyiasat perkara itu.

Jepun mengeluarkan amaran itu kepada rakyatnya berikutan laporan yang dikeluarkan oleh Site Intelligence Group, sebuah badan di Amerika Syarikat yang memantau komunikasi kalangan extremis di dalam talian.

Ketua Cawangan Anti-Pengganas Polis Diraja Malaysia Normah Ishak berkata bahawa amaran yang dikeluarkan Jepun itu hanyalah perkara rutin.

"Situasi keselamatan di Malaysia adalah terkawal dan pihak keselamatan juga senantiasa dalam keadaan siap siaga."

Indonesia pula menyatakan bahawa ia akan menyiasat maklumat itu, yang terkandung dalam pemakluman Jepun kepada rakyatnya yang berada di rantau itu.



KEMASKINI pada 8 malam waktu tempatan ETS - 2021-09-13

Pihak berkuasa Thai dan Malaysia pada Isnin menyatakan keadaan di negara masing-masing adalah selamat susulan amaran Jepun mengenai kemungkinan berlakunya serangan bom di Asia Tenggara, manakala polis anti-keganasan Indonesia berkata ia akan menyiasat perkara itu.

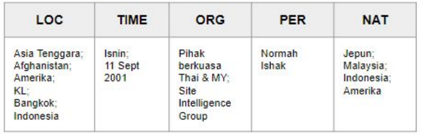
Jepun mengeluarkan amaran itu kepada rakyatnya berikutan laporan yang dikeluarkan oleh Site Intelligence Group, sebuah badan di Amerika Syarikat yang memantau komunikasi kalangan ekstremis di dalam talian. Badan itu mendakwa kumpulan pro Negara Islam di Afghanistan telah mengisytiharkan bahawa ini adalah "masanya untuk berperang" terutamanya di enam negara di Asia Tenggara.

Ketua Cawangan Anti-Pengganas Polis Diraja Malaysia Normah Ishak berkata bahawa amaran yang dikeluarkan Jepun itu hanyalah perkara rutin.

"Tidak ada maklumat kukuh mengenai ancaman serangan di negara kami. Rakan sejawatan kami di Jepun berkata bahawa ia adalah amalan biasa mereka untuk mengeluarkan amaran umum pada ulang tahun kejadian 9/11 selain perkembangan yang berlaku di Afghanistan," kata Normah kepada BenarNews, dengan merujuk kepada kejadian serangan maut pada 11 Sept 2001 di AS.

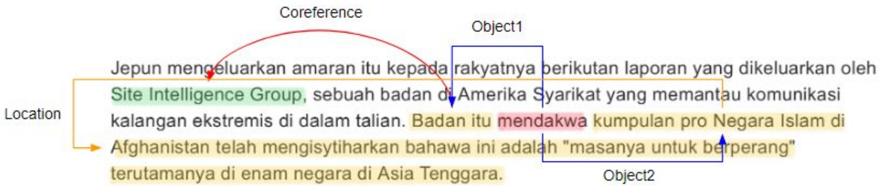
"Situasi keselamatan di Malaysia adalah terkawal dan pihak keselamatan juga sentiasa dalam keadaan siap siaga."

Kementerian Luar Malaysia berkata bahawa pemakluman Jepun itu telah dikongsi dengan



LOC	TIME	ORG	PER	NAT

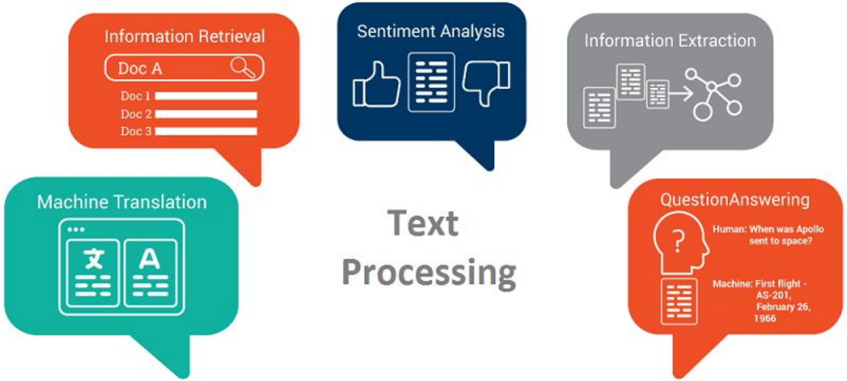
Asia Tenggara;
Afghanistan;
Amerika;
KL;
Bangkok;
Indonesia
	Isnin;
	11 Sept 2001
		Pihak berkuasa Thai & MY;
		Site Intelligence Group
			Normah Ishak	Jepun;
				Malaysia;
				Indonesia;
				Amerika



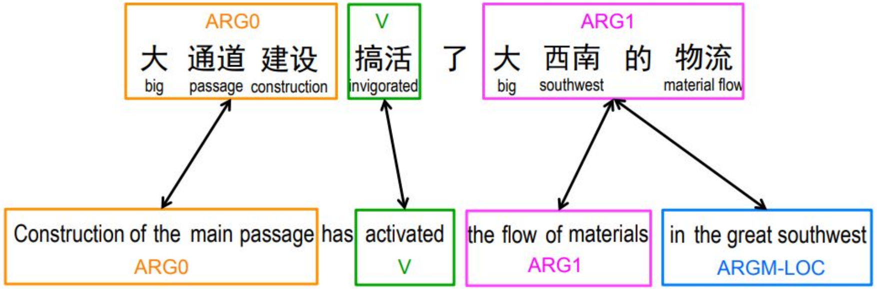
Jepun mengeluarkan amaran itu kepada rakyatnya berikutan laporan yang dikeluarkan oleh Site Intelligence Group, sebuah badan di Amerika Syarikat yang memantau komunikasi kalangan ekstremis di dalam talian. Badan itu mendakwa kumpulan pro Negara Islam di Afghanistan telah mengisytiharkan bahawa ini adalah "masanya untuk berperang" terutamanya di enam negara di Asia Tenggara.

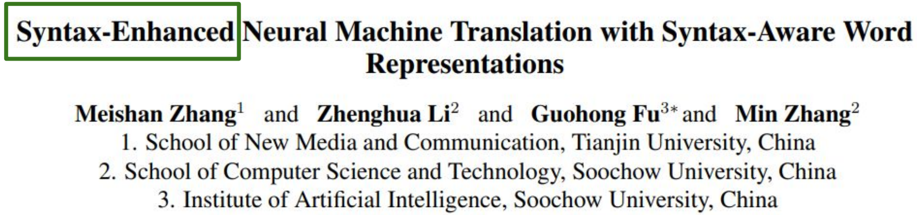
1 Direct Usage - Information Extraction (NER, Coreference Resolution, Semantic Role Labelling)

Impact - Trained Models



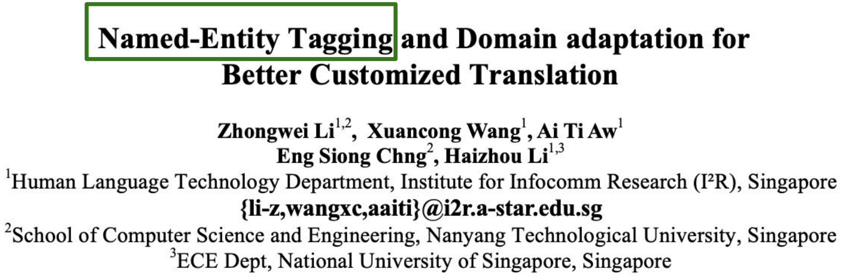
2 Feature Engineering for Downstream Tasks





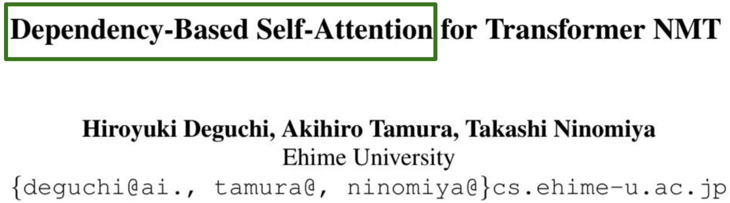
Syntax-Enhanced Neural Machine Translation with Syntax-Aware Word Representations

Meishan Zhang[1] and Zhenghua Li[2] and Guohong Fu[3]* and Min Zhang[2]
1. School of New Media and Communication, Tianjin University, China
2. School of Computer Science and Technology, Soochow University, China
3. Institute of Artificial Intelligence, Soochow University, China



Named-Entity Tagging and Domain adaptation for Better Customized Translation

Zhongwei Li[1,2], Xuancong Wang[1], Ai Ti Aw[1]
Eng Siong Chng[2], Haizhou Li[1,3]
[1]Human Language Technology Department, Institute for Infocomm Research (I²R), Singapore
{li-z,wangxc,aaiti}@i2r.a-star.edu.sg
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]ECE Dept, National University of Singapore, Singapore



Dependency-Based Self-Attention for Transformer NMT

Hiroyuki Deguchi, Akihiro Tamura, Takashi Ninomiya
Ehime University
{deguchi@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp

Outcome

Datasets for Public Use

Language	Dataset	Task	Size	Annotation	Paper
Indonesian	**COIN**	Coreference Resolution	2500 paragraphs, 730k tokens, 74k mentions	Complete	Under Review
	ICON	Constituency Parsing	10k sentences	Complete	Awaiting Review
		Semantic Role Labeling	10k sentences	Complete	About to begin
Thai		Coreference Resolution	30k sentences	In Progress	
		Dependency Parsing	30k sentences	In Progress	
		Semantic Role Labeling	30k sentences	In Progress	
Tamil		POS Tagging	10k sentences	In Progress	In Progress
		Named Entity Recognition	10k sentences	In Progress	In Progress
		Morphological Features	10k sentences	In Progress	In Progress
		Dependency Parsing	10k sentences	In Progress	In Progress

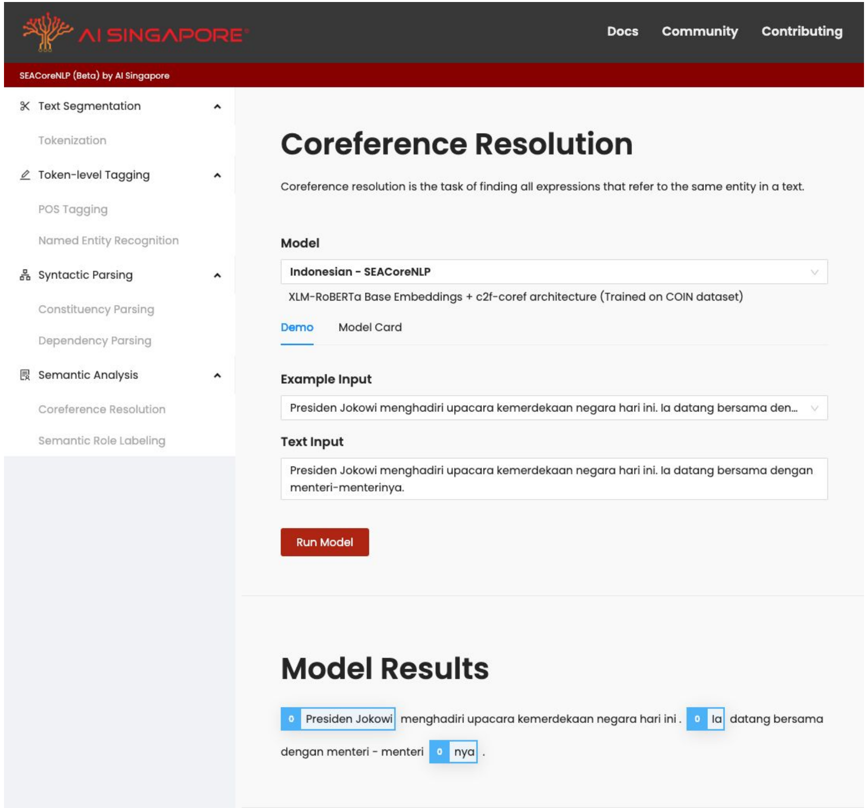
Outcome

More intra-regional knowledge transfer and collaboration

- Experience in linguistic data annotation
  - Guideline formulation
  - Tools and processes
- Training of regional multilingual language models
  - Austroasiatic/Tai-Kadai/MSEA Models
- Experience in NLP for low-resource languages

Repository of resources for SEA NLP

- Python package consolidating models for SEA Core NLP
- Documentation website consolidating research, datasets, and information about SEA NLP
- Demo website to showcase capabilities of models to attract the industry



Coreference Resolution

Coreference resolution is the task of finding all expressions that refer to the same entity in a text.

Model

Indonesian - SEACoreNLP

XLM-RoBERTa Base Embeddings + c2f-coref architecture (Trained on COIN dataset)

Example Input

Presiden Jokowi menghadiri upacara kemerdekaan negara hari ini. Ia datang bersama den...

Text Input

Presiden Jokowi menghadiri upacara kemerdekaan negara hari ini. Ia datang bersama dengan menteri-menterinya.

Model Results

0 Presiden Jokowi menghadiri upacara kemerdekaan negara hari ini . 0 Ia datang bersama dengan menteri - menteri 0 nya .

https://seacorenlp.aisingapore.net/

Conclusion

Goal

- To improve NLP capabilities in SEA (ASEAN) languages

Solution/ Impact

- Build high-quality benchmark datasets for
  - Training, evaluating and probing models
  - Catalyzing research and development
- Consolidate existing models/tools and fill in the gaps
  - Direct use for applications like information extraction
  - Indirect use (feature engineering for downstream tasks)
- Establish a regional coalition to improve knowledge flow and resource sharing

Accomplishments

- Published Python package (seacorenlp)
- Demo website for showcasing SEA NLP capabilities
- Documentation website for consolidating resources
- Indonesian dataset annotation completed
  - Constituency Parsing
  - Coreference Resolution
  - Semantic Role Labelling
- Two papers (Indo CP and Coref) under review
- Established preliminary regional network
- Worked with industry to accumulate use cases

Future works

- Scaling up operations to more languages and tasks and larger data volumes
- Building more monolingual/multilingual language models for ASEAN languages
- Grow the regional network to more countries and organizations

Acknowledgements

Indonesia

- Assoc. Prof. Dr. Ayu Purwarianti (ITB/Prosa.ai)
- Dea Adhista and her team (Prosa.ai)

Thailand

- Asst. Prof. Dr. Attapol Rutherford (CU)
- Assoc. Prof. Dr. Sarana Nutanong (VISTEC)
- Charin

Vietnam

- Assoc. Prof. Dr. Nguyen Phuong Thai (VNU)

The Philippines

- Prof. Dr. Regina Estuar (Ateneo)
- Dr. Rachel Roxas

Singapore

- Assoc. Prof. Dr. Titima Suthiwan (NUS)
- Arie Pratama Sutiono

India

- Asst. Prof. Dr. Rajiv Ratn Shah (IIIT-Delhi)

Sri Lanka

- Dr. Kengatharaiyer Sarveswaran (Jaffna)

Annotation Platform

- Datasaur (https://datasaur.ai)