Re-targetable OCR and Cambodian Gazetteer

Mudit Agrawal
LAMP LAB
University of Maryland, College Park

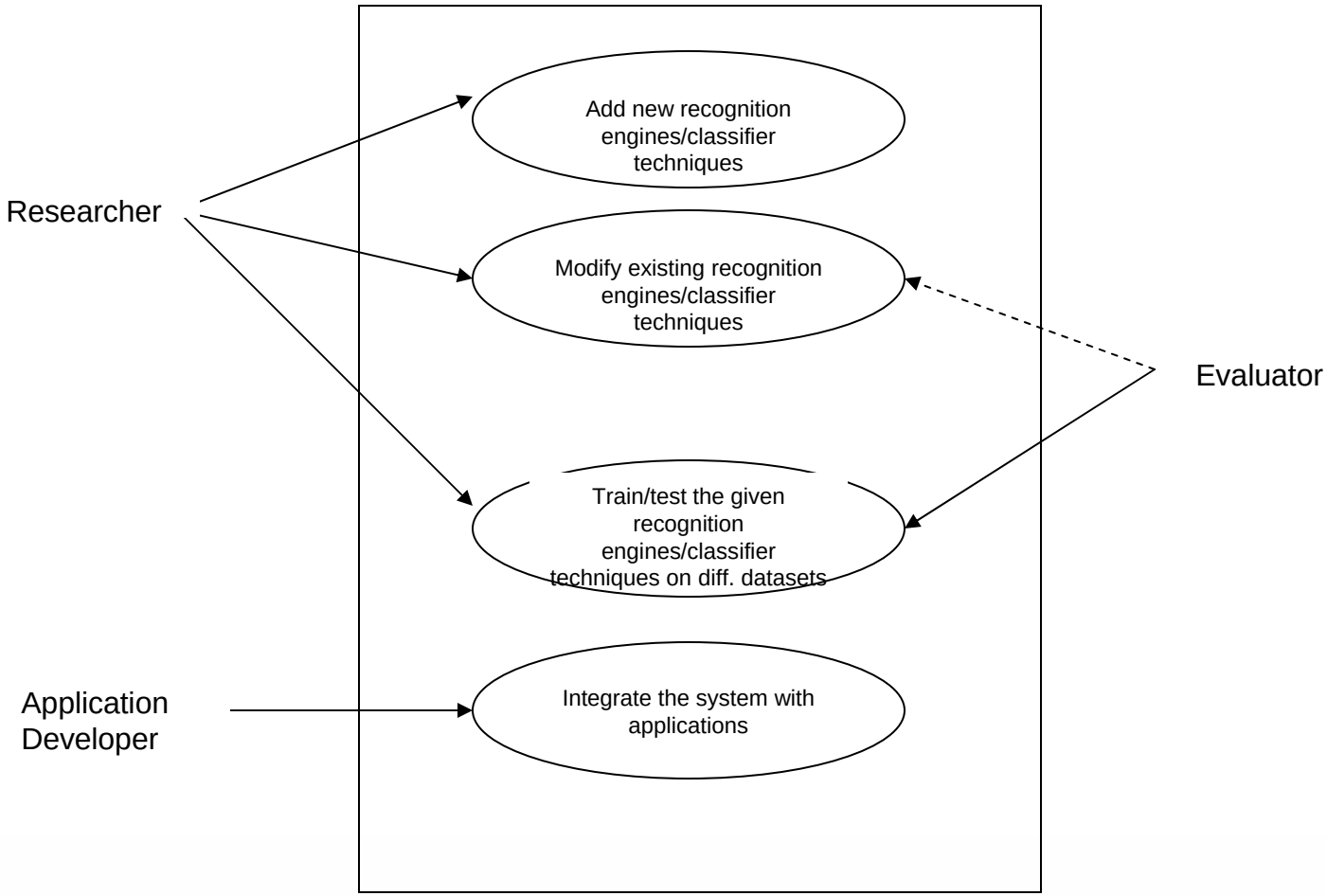
Need

- An adaptive system which can train/learn a new script
- Optimized for noise and font characteristics of target document
- Must requirement
  - minimal user interaction and
  - a minimal number of samples

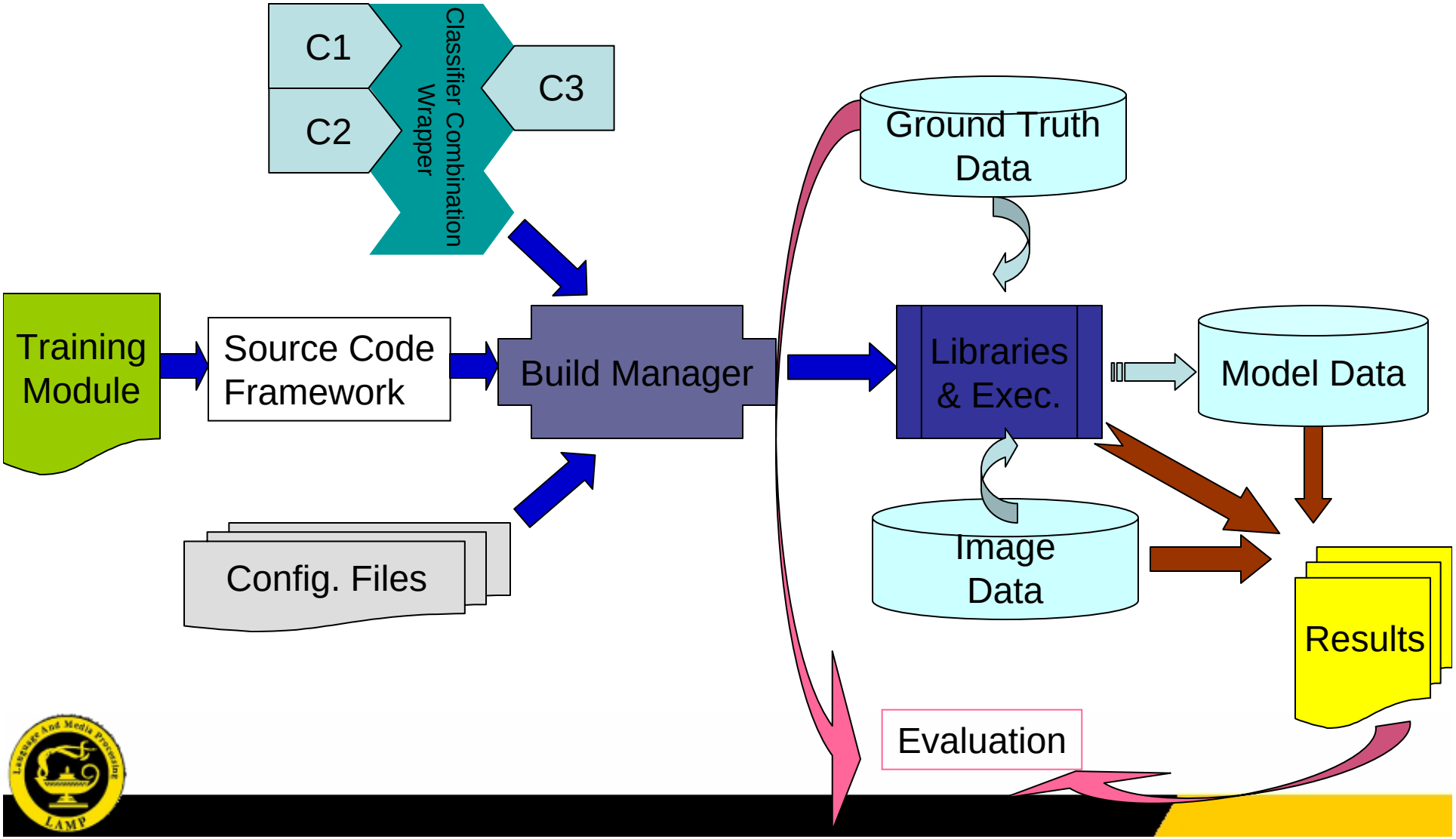
Objective

- To design a script-independent printed character recognition system which should provide the flexibility of
  - Training and testing using different datasets
  - Using any new script
  - Plugging new recognition algorithms/classifiers

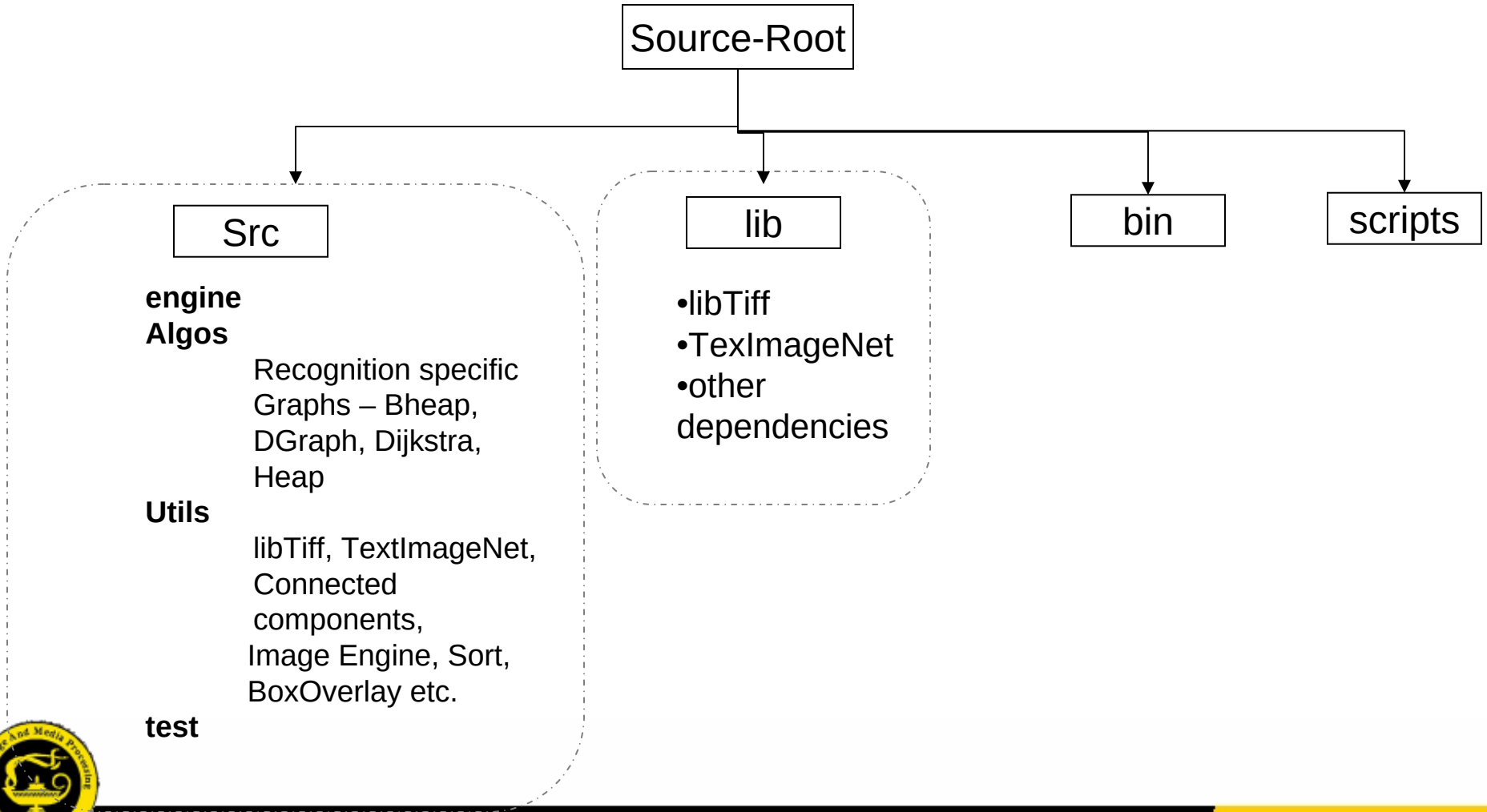
Use-case diagram



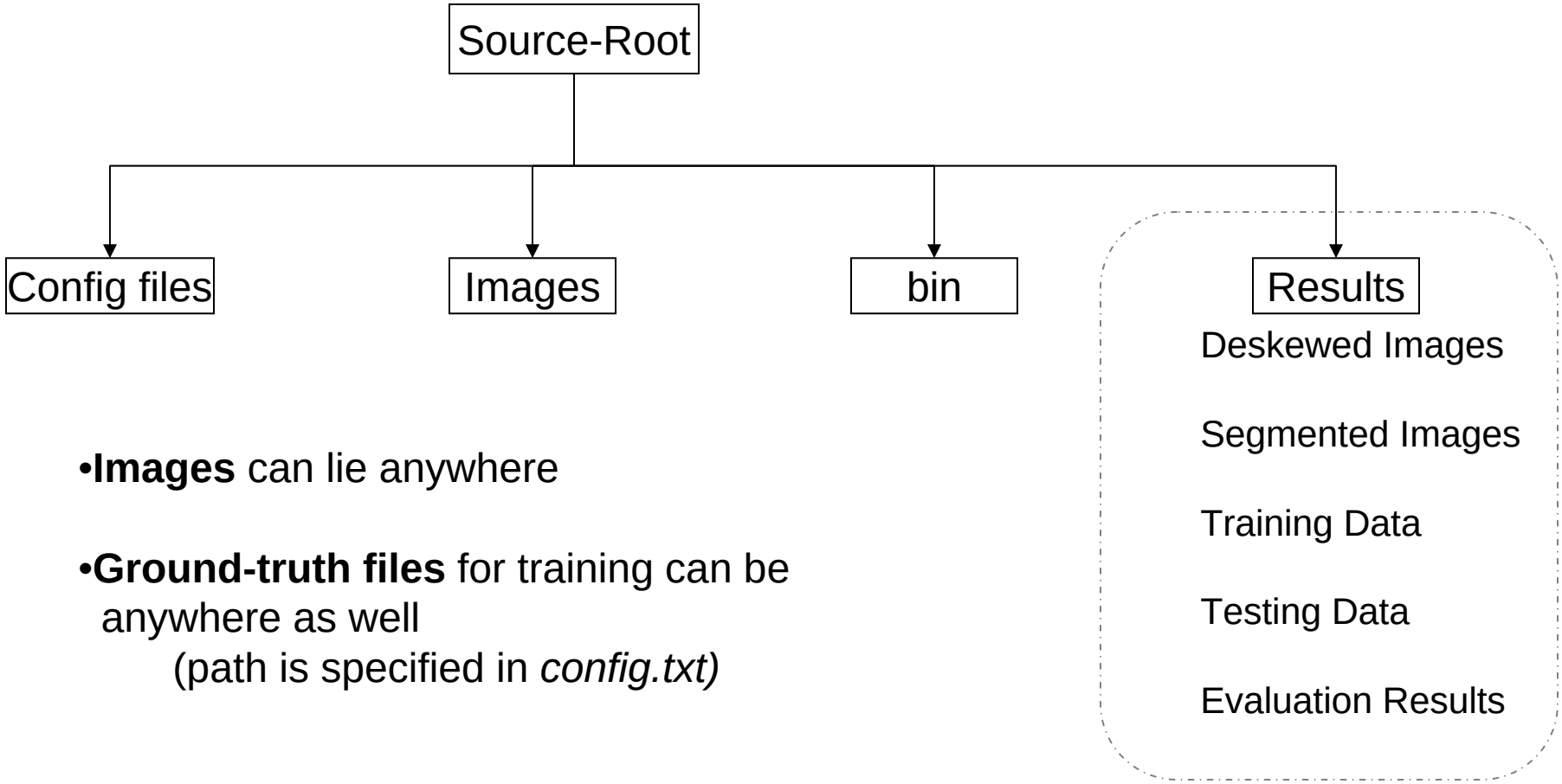
System Overview



Organization



System



- **Images** can lie anywhere
- **Ground-truth files** for training can be anywhere as well
  (path is specified in *config.txt)*

Training Task

- Training is done by
  - Aligning the ground-truth text with the image-document
    - Alignment is done in the order of line, word and character mapping
    - Using approximate font style to do char-segmentation

Line Mapping

គេតែងទទួលស្គាល់ជាទូទៅថាផ្លូវថ្នល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍ ហើយផ្លូវថ្នល់ក៏ជួយកាត់
គេតែងទទួលស្គាល់ជាទូទៅថាផ្លូវថ្នល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍ ហើយផ្លូវថ្នល់ក៏ជួយកាត់

បន្ថយភាពក្រីក្រផងដែរ។ ទោះជាយ៉ាងនេះក្តី ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់
បន្ថយភាពក្រីក្រផងដែរ។ ទោះជាយ៉ាងនេះក្តី ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់

ផ្លូវថ្នល់ គេចាំបាច់ត្រូវធ្វើការថែទាំផ្លូវថ្នល់អោយបានត្រឹមត្រូវ។ នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើន
ផ្លូវថ្នល់ គេចាំបាច់ត្រូវធ្វើការថែទាំផ្លូវថ្នល់អោយបានត្រឹមត្រូវ។ នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើន

ដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី ស្តង់ដារនៃការថែទាំមានកិរិតខ្សោយ ។ ជាញឹកញយ
ដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី ស្តង់ដារនៃការថែទាំមានកិរិតខ្សោយ ។ ជាញឹកញយ

មូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការថែទាំផ្លូវមិនបានគ្រប់គ្រាន់។
មូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការថែទាំផ្លូវមិនបានគ្រប់គ្រាន់។

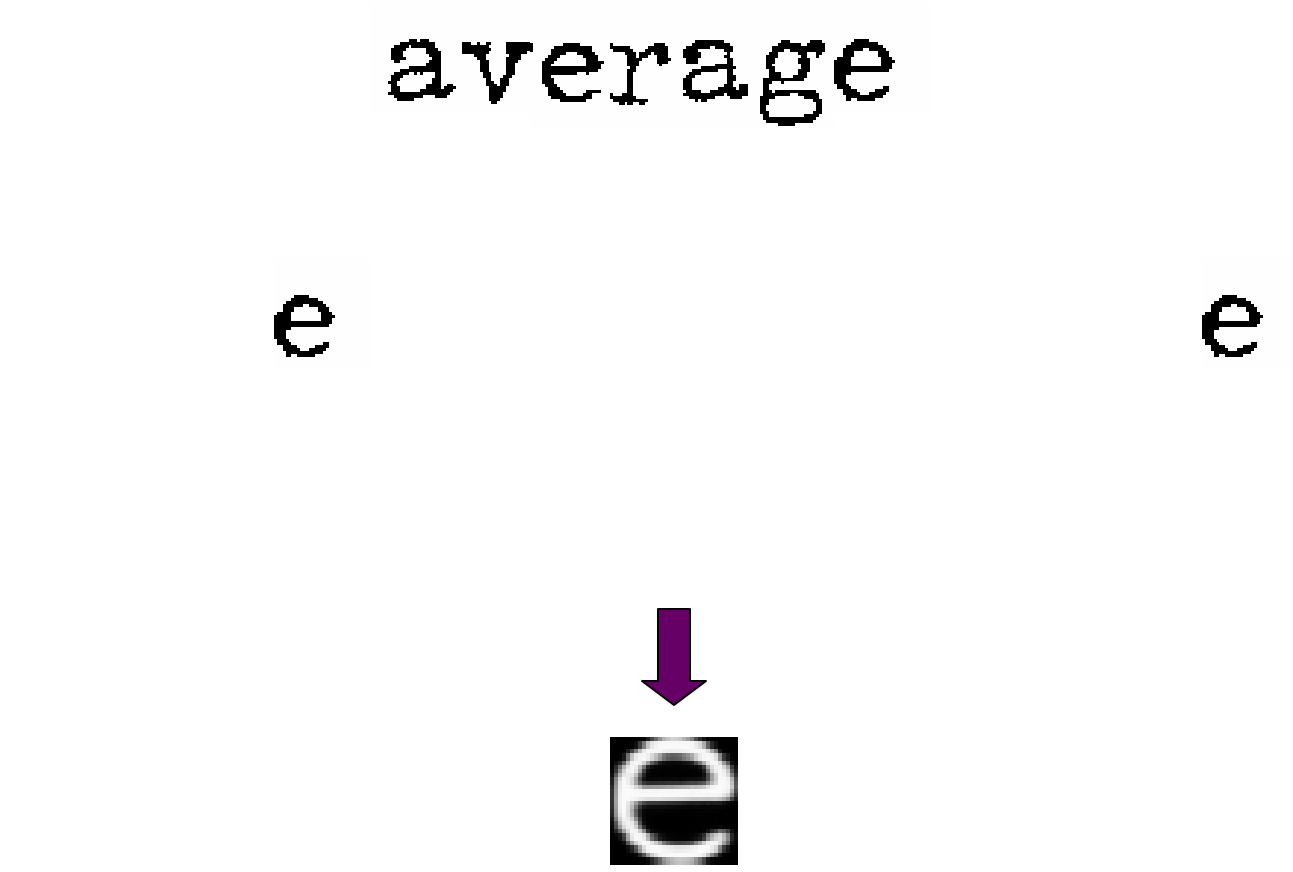
របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគារ
របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគារ

អភិវឌ្ឍន៍អាស៊ី លេខ អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធីចំពោះផ្លូវថ្នល់ ដើម្បីពិនិត្យមើល
អភិវឌ្ឍន៍អាស៊ី លេខ អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធីចំពោះផ្លូវថ្នល់ ដើម្បីពិនិត្យមើល

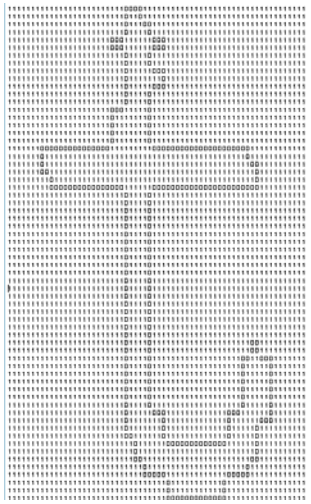
Word Mapping

គេតែងទទួលស្គាល់ជាទូទៅថាផ្ទះវិថ្ពល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍	គេតែងទទួលស្គាល់ជាទូទៅថាផ្ទះវិថ្ពល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍
ហើយផ្ទះវិថ្ពល់ក៏ជួយកាត់	ហើយផ្ទះវិថ្ពល់ក៏ជួយកាត់
បន្ថយភាពក្រីក្រផងដែរ ។	បន្ថយភាពក្រីក្រផងដែរ ។
ទោះជាយ៉ាងនេះក្តី	ទោះជាយ៉ាងនេះក្តី
ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់	ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់
ផ្ទះវិថ្ពល់	ផ្ទះវិថ្ពល់
គេចាំបាច់ត្រូវធ្វើការថែទាំផ្ទះវិថ្ពល់អោយបានត្រឹមត្រូវ ។	គេចាំបាច់ត្រូវធ្វើការថែទាំផ្ទះវិថ្ពល់អោយបានត្រឹមត្រូវ ។
នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើន	នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើន
ដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី	ដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី
ស្តង់ដារនៃការថែទាំមានកិរិតខ្សោយ	ស្តង់ដារនៃការថែទាំមានកិរិតខ្សោយ
។	។
ជាញឹកញាយ	ជាញឹកញាយ
មូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការថែទាំផ្ទះវិមិនបានគ្រប់គ្រាន់ ។	មូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការថែទាំផ្ទះវិមិនបានគ្រប់គ្រាន់ ។
របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគារ	របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគា
អភិវឌ្ឍន៍អាស៊ី	អភិវឌ្ឍន៍អាស៊ី
លេខ	លេខ
អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធីចំពោះផ្ទះវិថ្ពល់	អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធីចំពោះផ្ទះវិថ្ពល់
ដើម្បីពិនិត្យមើល	ដើម្បីពិនិត្យមើល
បញ្ហាផ្តល់មូលនិធិលើផ្នែកផ្ទះវិថ្ពល់	បញ្ហាផ្តល់មូលនិធិលើផ្នែកផ្ទះវិថ្ពល់

Featurization: *Template Matching*

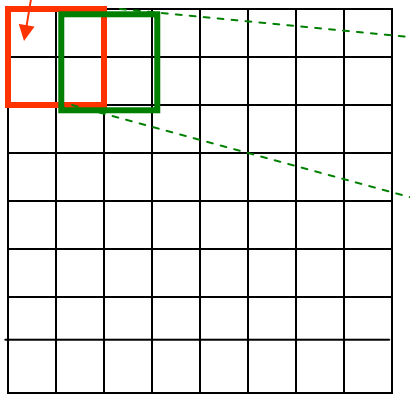


Featurization: *Directional*

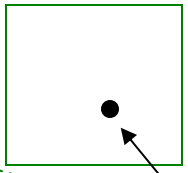


Normalization and contour extraction

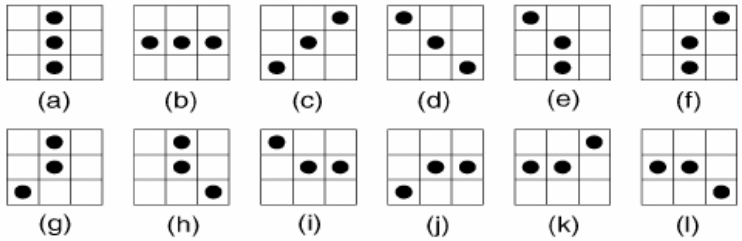
16 X 16 pixels



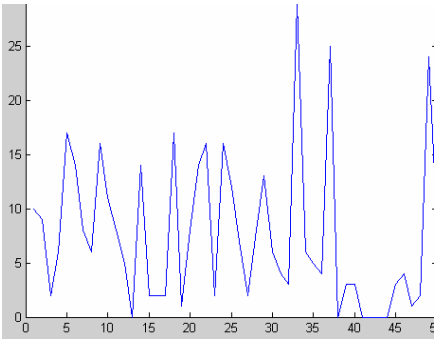
Dividing into 7 X 7 blocks of 16 X 16 pixel



For every pixel, look at its neighbors and maintain a record of Vertical, horizontal, two oblique lines using



Aggregate records For every block

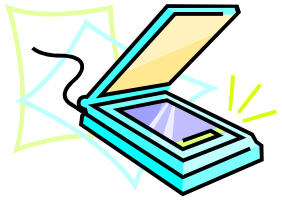


Feature vector

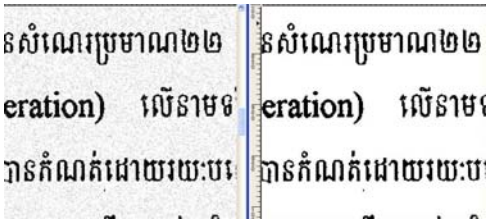
Classifier combination

- *Template Matching*: awarding probabilities where template pixel matches with the test char pixel and penalizing otherwise
- *CityBlockDistance with Deviation*: $$d_{CBDD}(\mathbf{v}) = \sum_{j=1}^{n} \max\left\{0, \left|v_j - \mu_j\right| - \theta \cdot s_j\right\},$$
- A generic framework to combine different classifiers
- This forms the backbone of experiments to test
  - The validity of classifier(s) on a given dataset
  - Dependence of the script on that classifier combination
  - Combination scheme – depending on the evaluation results
  - *Coming up with best suited combinations specific to the script under study*

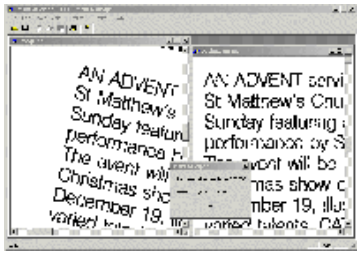
Work-flow



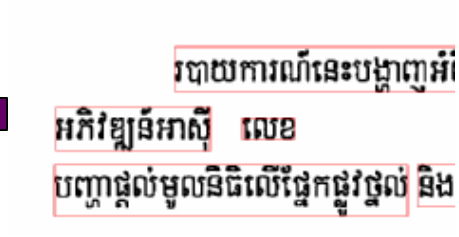
Scanning



Noise Removal



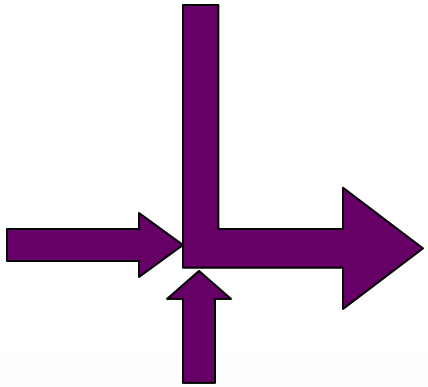
Skew Detection



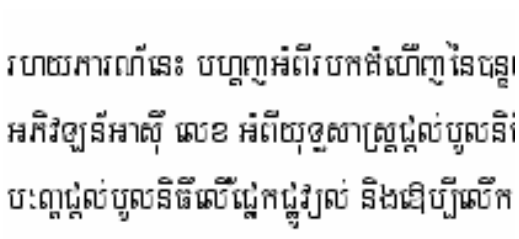
Word Extraction

Char Extraction

Templates



Recognizer



Unicode Text

Testing

- Testing should be possible on
  - Image Documents with several formats (like tif, png, bmp etc.)
- Plugging the testing results with evaluation tools
- A feedback training system using the evaluation tool
  - To know certain special features of the new script under study
  - To tune the system for that script

Region & Line Segmentation

គេតែងទទួលស្គាល់ជាទូទៅថាផ្លូវថ្នល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍ ហើយផ្លូវថ្នល់ក៏ជួយកាត់បន្ថយភាពក្រីក្រផងដែរ។ ទោះជាយ៉ាងនេះក្តី ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់ផ្លូវថ្នល់ គេចាំបាច់ត្រូវធ្វើការថែទាំផ្លូវថ្នល់អោយបានត្រឹមត្រូវ។ នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើនដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី ស្តង់ដារនៃការថែទាំមានកិរិតខ្សោយ ។ ជាញឹកញយមូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការថែទាំផ្លូវមិនបានត្រប់ត្រាន់។

គេតែងទទួលស្គាល់ជាទូទៅថាផ្លូវថ្នល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍ ហើយផ្លូវថ្នល់ក៏ជួយកាត់បន្ថយភាពក្រីក្រផងដែរ។ ទោះជាយ៉ាងនេះក្តី ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់ផ្លូវថ្នល់ គេចាំបាច់ត្រូវធ្វើការថែទាំផ្លូវថ្នល់អោយបានត្រឹមត្រូវ។ នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើនដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី ស្តង់ដារនៃការថែទាំមានកិរិតខ្សោយ ។ ជាញឹកញយមូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការថែទាំផ្លូវមិនបានត្រប់ត្រាន់។

របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគារអភិវឌ្ឍន៍អាស៊ី លេខ អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធិចំពោះផ្លូវថ្នល់ ដើម្បីពិនិត្យមើលបញ្ហាផ្តល់មូលនិធិលើផ្នែកផ្លូវថ្នល់ និងដើម្បីលើកសំណើដំណោះស្រាយ។

របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគារអភិវឌ្ឍន៍អាស៊ី លេខ អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធិចំពោះផ្លូវថ្នល់ ដើម្បីពិនិត្យមើលបញ្ហាផ្តល់មូលនិធិលើផ្នែកផ្លូវថ្នល់ និងដើម្បីលើកសំណើដំណោះស្រាយ។

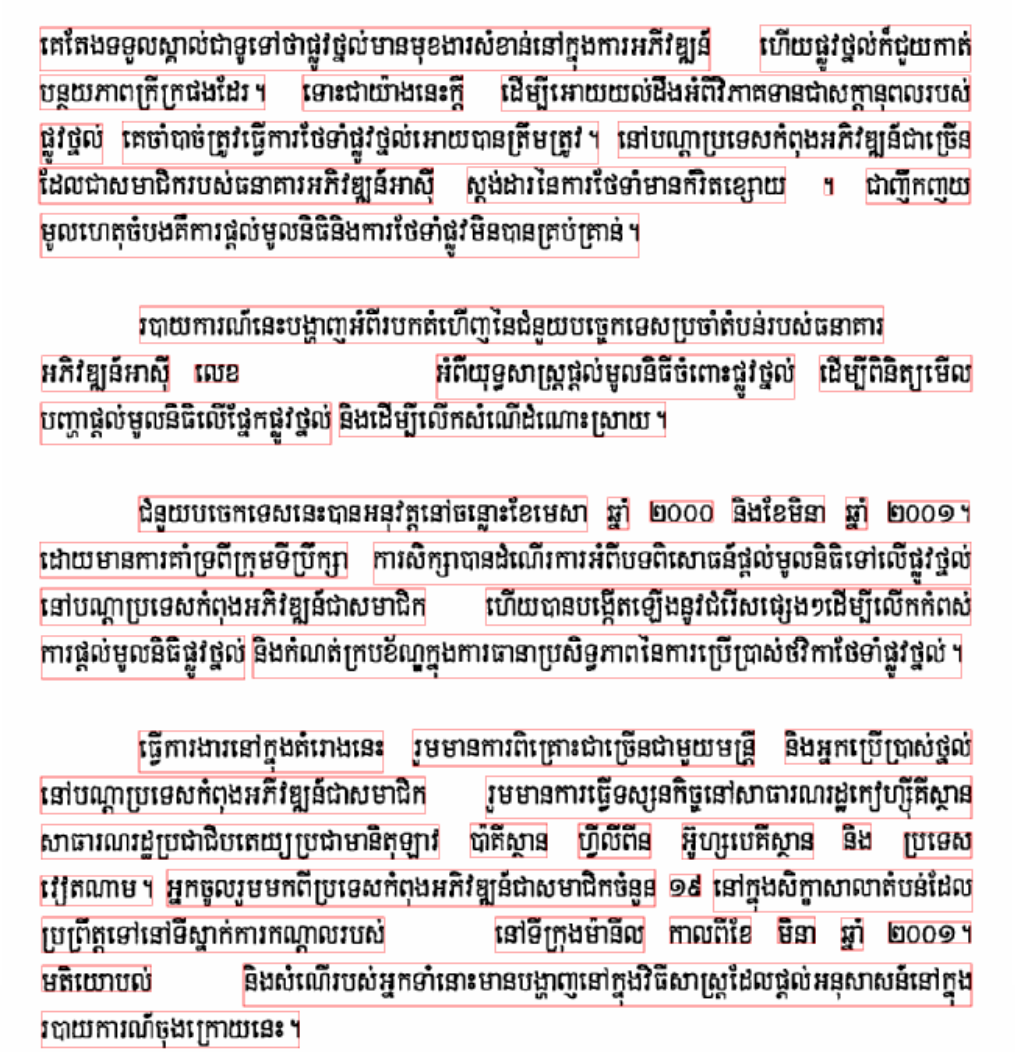
ជំនួយបចេកទេសនេះបានអនុវត្តនៅចន្លោះខែមេសា ឆ្នាំ ២០០០ និងខែមិនា ឆ្នាំ ២០០១។ ដោយមានការគាំទ្រពីក្រុមទីប្រឹក្សា ការសិក្សាបានដំណើរការអំពីបទពិសោធន៍ផ្តល់មូលនិធិទៅលើផ្លូវថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក ហើយបានបង្កើតឡើងនូវជំពើសផ្សេង១ដើម្បីលើកកំពស់ការផ្តល់មូលនិធិផ្លូវថ្នល់ និងកំណត់ក្របខ័ណ្ឌក្នុងការធានាប្រសិទ្ធភាពនៃការប្រើប្រាស់ថវិកាថែទាំផ្លូវថ្នល់។

ជំនួយបចេកទេសនេះបានអនុវត្តនៅចន្លោះខែមេសា ឆ្នាំ ២០០០ និងខែមិនា ឆ្នាំ ២០០១។ ដោយមានការគាំទ្រពីក្រុមទីប្រឹក្សា ការសិក្សាបានដំណើរការអំពីបទពិសោធន៍ផ្តល់មូលនិធិទៅលើផ្លូវថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក ហើយបានបង្កើតឡើងនូវជំពើសផ្សេង១ដើម្បីលើកកំពស់ការផ្តល់មូលនិធិផ្លូវថ្នល់ និងកំណត់ក្របខ័ណ្ឌក្នុងការធានាប្រសិទ្ធភាពនៃការប្រើប្រាស់ថវិកាថែទាំផ្លូវថ្នល់។

ធ្វើការងារនៅក្នុងតំរោងនេះ រួមមានការពិគ្រោះជាច្រើនជាមួយមន្ត្រី និងអ្នកប្រើប្រាស់ថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក រួមមានការធ្វើទស្សនកិច្ចនៅសាធារណរដ្ឋកៀហ្ស៊ីគីស្ថាន សាធារណរដ្ឋប្រជាជិបតេយ្យប្រជាមានិតមុឡាវ ប៉ាគីស្ថាន ហ្វីលីពីន អ៊ូហ្សបេគីស្ថាន និង ប្រទេសវៀតណាម។ អ្នកចូលរួមមកពីប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិកចំនួន ១៩ នៅក្នុងសិក្ខាសាលាតំបន់ដែលប្រព្រឹត្តទៅនៅទីស្នាក់ការកណ្តាលរបស់ នៅទីក្រុងម៉ានីល កាលពីខែ មិនា ឆ្នាំ ២០០១។ មតិយោបល់ និងសំណើរបស់អ្នកទាំនោះមានបង្ហាញនៅក្នុងវិធីសាស្ត្រដែលផ្តល់អនុសាសន៍នៅក្នុងរបាយការណ៍ចុងក្រោយនេះ។

ធ្វើការងារនៅក្នុងតំរោងនេះ រួមមានការពិគ្រោះជាច្រើនជាមួយមន្ត្រី និងអ្នកប្រើប្រាស់ថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក រួមមានការធ្វើទស្សនកិច្ចនៅសាធារណរដ្ឋកៀហ្ស៊ីគីស្ថាន សាធារណរដ្ឋប្រជាជិបតេយ្យប្រជាមានិតមុឡាវ ប៉ាគីស្ថាន ហ្វីលីពីន អ៊ូហ្សបេគីស្ថាន និង ប្រទេសវៀតណាម។ អ្នកចូលរួមមកពីប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិកចំនួន ១៩ នៅក្នុងសិក្ខាសាលាតំបន់ដែលប្រព្រឹត្តទៅនៅទីស្នាក់ការកណ្តាលរបស់ នៅទីក្រុងម៉ានីល កាលពីខែ មិនា ឆ្នាំ ២០០១។ មតិយោបល់ និងសំណើរបស់អ្នកទាំនោះមានបង្ហាញនៅក្នុងវិធីសាស្ត្រដែលផ្តល់អនុសាសន៍នៅក្នុងរបាយការណ៍ចុងក្រោយនេះ។

Word Segmentation



Syllable and Character Seg

តែតែងទទួលស្គាល់ជាទូទៅថាផ្លូវថ្នល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍ ហើយផ្លូវថ្នល់ក៏ជួយកាត់បន្ថយភាពក្រីក្រជងដែរ។ នោះជាយ៉ាងនេះក្តី ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់ផ្លូវថ្នល់ គេចាំបាច់ត្រូវធ្វើការវិចារណ៍ផ្លូវថ្នល់អោយបានត្រឹមត្រូវ។ នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើនដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី ស្តង់ដារនៃការវិចារណ៍មានកិរិតខ្សោយ ។ ជាញឹកញយមូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការវិចារណ៍ផ្លូវមិនបានត្រប់ត្រាន់។

របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគារអភិវឌ្ឍន៍អាស៊ី លេខ អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធិចំពោះផ្លូវថ្នល់ ដើម្បីពិនិត្យមើលបញ្ហាផ្តល់មូលនិធិលើផ្នែកផ្លូវថ្នល់ និងដើម្បីលើកសំណើដំណោះស្រាយ។

ជំនួយបច្ចេកទេសនេះបានអនុវត្តនៅចន្លោះខែមេសា ឆ្នាំ ២០០០ និងខែមីនា ឆ្នាំ ២០០១។ ដោយមានការគាំទ្រពីក្រុមមន្ត្រីប្រឹក្សា ការសិក្សាបានដឹងលើរការអំពីបទពិសោធន៍ផ្តល់មូលនិធិទៅលើផ្លូវថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក ហើយបានបង្កើតឡើងនូវជំរើសផ្សេង១ដើម្បីលើកកំពស់ការផ្តល់មូលនិធិផ្លូវថ្នល់ និងកំណត់ក្របខ័ណ្ឌក្នុងការធានាប្រសិទ្ធភាពនៃការប្រើប្រាស់ថវិការថែទាំផ្លូវថ្នល់។

ធ្វើការងាររនៅក្នុងតំរោងនេះ រួមមានការពិគ្រោះជាច្រើនជាមួយមន្ត្រី និងអ្នកប្រើប្រាស់ថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក រួមមានការធ្វើទស្សនកិច្ចនៅសាធារណរដ្ឋក្បៀហ្ស៊ីគីស្ថាន សាធារណរដ្ឋប្រជាជិបតេយ្យប្រជាមានិតឡាវ ម៉ាគីស្ថាន ហ្វីលីពីន អ៊ូបេគីស្ថាន និង ប្រទេសវៀតណាម។ អ្នកចូលរួមមកពីប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិកចំនួន ១៩ នៅក្នុងសិក្ខាសាលាតំបន់ដែលប្រព្រឹត្តទៅនៅទីស្នាក់ការកណ្តាលរបស់ នៅទីក្រុងម៉ានីល កាលពីខែ មីនា ឆ្នាំ ២០០១។ មតិយោបល់ និងសំណើរបស់អ្នកទាំនោះមានបង្ហាញនៅក្នុងវិធីសាស្ត្រដែលផ្តល់អនុសាសន៍នៅក្នុងរបាយការណ៍ចុងក្រោយនេះ។

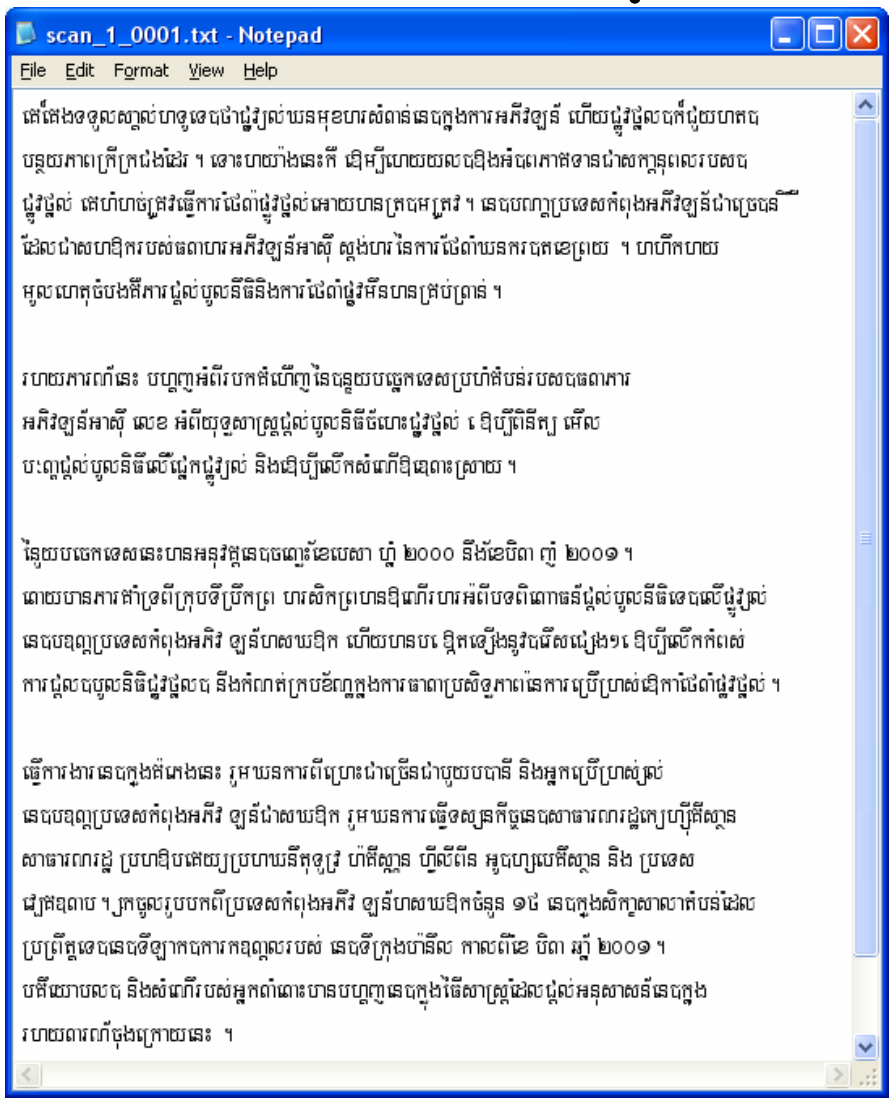
តែតែងទទួលស្គាល់ជាទូទៅថាផ្លូវថ្នល់មានមុខងារសំខាន់នៅក្នុងការអភិវឌ្ឍន៍ ហើយផ្លូវថ្នល់ក៏ជួយកាត់បន្ថយភាពក្រីក្រជងដែរ។ នោះជាយ៉ាងនេះក្តី ដើម្បីអោយយល់ដឹងអំពីវិភាគទានជាសក្តានុពលរបស់ផ្លូវថ្នល់ គេចាំបាច់ត្រូវធ្វើការវិចារណ៍ផ្លូវថ្នល់អោយបានត្រឹមត្រូវ។ នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រើនដែលជាសមាជិករបស់ធនាគារអភិវឌ្ឍន៍អាស៊ី ស្តង់ដារនៃការវិចារណ៍មានកិរិតខ្សោយ ។ ជាញឹកញយមូលហេតុចំបងគឺការផ្តល់មូលនិធិនិងការវិចារណ៍ផ្លូវមិនបានត្រប់ត្រាន់។

របាយការណ៍នេះបង្ហាញអំពីរបកគំហើញនៃជំនួយបច្ចេកទេសប្រចាំតំបន់របស់ធនាគារអភិវឌ្ឍន៍អាស៊ី លេខ អំពីយុទ្ធសាស្ត្រផ្តល់មូលនិធិចំពោះផ្លូវថ្នល់ ដើម្បីពិនិត្យមើលបញ្ហាផ្តល់មូលនិធិលើផ្នែកផ្លូវថ្នល់ និងដើម្បីលើកសំណើដំណោះស្រាយ។

ជំនួយបច្ចេកទេសនេះបានអនុវត្តនៅចន្លោះខែមេសា ឆ្នាំ ២០០០ និងខែមីនា ឆ្នាំ ២០០១។ ដោយមានការគាំទ្រពីក្រុមមន្ត្រីប្រឹក្សា ការសិក្សាបានដឹងលើរការអំពីបទពិសោធន៍ផ្តល់មូលនិធិទៅលើផ្លូវថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក ហើយបានបង្កើតឡើងនូវជំរើសផ្សេង១ដើម្បីលើកកំពស់ការផ្តល់មូលនិធិផ្លូវថ្នល់ និងកំណត់ក្របខ័ណ្ឌក្នុងការធានាប្រសិទ្ធភាពនៃការប្រើប្រាស់ថវិការថែទាំផ្លូវថ្នល់។

ធ្វើការងាររនៅក្នុងតំរោងនេះ រួមមានការពិគ្រោះជាច្រើនជាមួយមន្ត្រី និងអ្នកប្រើប្រាស់ថ្នល់នៅបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិក រួមមានការធ្វើទស្សនកិច្ចនៅសាធារណរដ្ឋក្បៀហ្ស៊ីគីស្ថាន សាធារណរដ្ឋប្រជាជិបតេយ្យប្រជាមានិតឡាវ ម៉ាគីស្ថាន ហ្វីលីពីន អ៊ូបេគីស្ថាន និង ប្រទេសវៀតណាម។ អ្នកចូលរួមមកពីប្រទេសកំពុងអភិវឌ្ឍន៍ជាសមាជិកចំនួន ១៩ នៅក្នុងសិក្ខាសាលាតំបន់ដែលប្រព្រឹត្តទៅនៅទីស្នាក់ការកណ្តាលរបស់ នៅទីក្រុងម៉ានីល កាលពីខែ មីនា ឆ្នាំ ២០០១។ មតិយោបល់ និងសំណើរបស់អ្នកទាំនោះមានបង្ហាញនៅក្នុងវិធីសាស្ត្រដែលផ្តល់អនុសាសន៍នៅក្នុងរបាយការណ៍ចុងក្រោយនេះ។

OCR Output


scan_1_0001.txt - Notepad

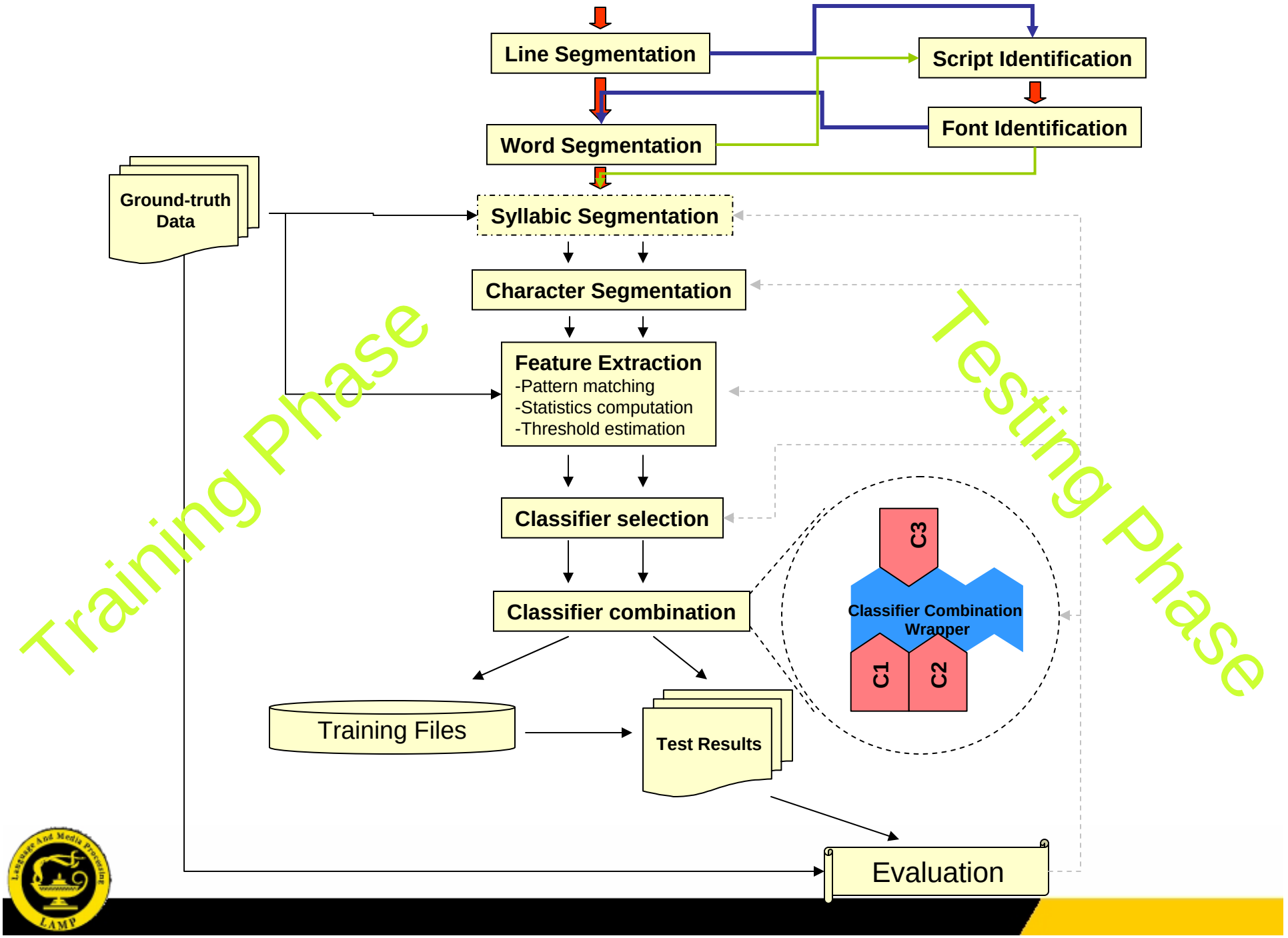
File Edit Format View Help

តើដែងទទួលស្គាល់ហេតុទេបបំាប្លែវ្យាល់ឃនអុខហារសំពាន់នេបក្តងការអភិវឌ្ឍន៍ ហើយប្លែវដ្ឋែលបក៏ជួយហាតប
បន្តយភាពក្រីក្រដែងដែរ ។ ទោះហយ៉ាងនេះកី ដើអ្បីហោយយលបឱ្យងអំបពោភាគទានជាសក្តនុពលរបសប
ប្លែវដ្ឋែល់ តេហំាហច់ត្រែវធ្វើការជែពាំប្លែវដ្ឋែល់អោយហានត្របអត្រែវ ។ នេបបណ្តាប្រទេសកំពុងអភិវឌ្ឍន៍ជាច្រេបនែ
ដែលជាសហគិករបស់ធពាហារអភិវឌ្ឍន៍អាស៊ី ស្តង់ហារនៃការជែពាំឃនករបតខេត្រយ ។ ហហើកហាយ
អូលហេតុចំបងគឺការប្តែល់បូលនីធិនិងការជែពាំប្លែវអីនហានត្រែប់ត្រាន់ ។

រហាយការណ៍នេះ បហ្តាញអំពីរបកគំហើញនៃបន្តយបច្ចេកទេសប្រហាំគំបន់របសបធពាការ
អភិវឌ្ឍន៍អាស៊ី ដេខ អំពីយុទ្ធសាស្ត្រប្តែល់បូលនិធីចំហោះប្លែវដ្ឋែល់ ដេីប្បីពិនិត្យ អើល
បៈល្អប្តែល់បូលនិធីដេីប្តើកប្លែវ្យាល់ និងដេីប្បីដេីកសំណេីរិដេពារស្រាយ ។

នៃយបចេកទេសនេះហានអនុវត្តនេបចន្លោះខែដេសា ហ្គ្គ ២០០០ នីងខែបិតា ក្ញ ២០០១ ។
ដោយហានការគាំទ្រពីក្រុបចីប្រឹកព្រ ហរសិកព្រហានឱ្យភីរហារអំពីបទពិដោធន៍ប្តែល់បូលនីធិទេបដេីប្លែវ្យាល់
នេបបឧុណ្ណប្រទេសកំពុងអភិវ ឡន៍ហាសយឱ្យក ហើយហានបដេ្តីតឡើងនូវបជីសជ្បែង១ដេីប្បីដេីកកំពស់
ការប្តែលបបូលនិធីប្លែវដ្ឋែលប នីងកំពាត់ក្របខ័ណ្ឌក្តងការធាពាប្រសិទ្ធភាពនែការប្រើប្រាស់ដេីការជែពាំប្លែវដ្ឋែល់ ។

ធ្វើការងារនេបក្នងតំហេងនេះ រូអយនការពិប្រេះជាច្រេីនជាបួយបហានី និងអ្នកប្រើប្រាស់ស្រប់
នេបបឧុណ្ណប្រទេសកំពុងអភិវ ឡន៍ជាសយឱ្យក រូអយនការធ្វើទស្សនកិច្ចនេបសាធារណរដ្ឋកេ្បហ្ប៊ីគីស្ថាន
សាធារណរដ្ឋ ប្រហាឱ្បងដេយ្យប្រហាយនីតុទ្ធ្រែ ហ៊ាគីស្ថាន ហ្ស៊ីលីពីន អូបហ្សេគីស្ថាន និង ប្រទេស
ជ្បែគឧពាប ។ ក្រចូលរូបបកពីប្រទេសកំពុងអភិវ ឡន៍ហាសយឱ្យកចំនួន ១៥ នេបក្នងសិក្ខាសាលាតំបន់ដែល
ប្រព្រឹត្តទេបនេបចីឡាកបការកឧុណ្ណលរបស់ នេបចីក្រុងប៉ានីល កាលពីដេ បិតា ឆ្ន្ល ២០០១ ។
បតីយោបលប និងសំណេីរបស់អ្នកពាំណោះហានបហ្តាញនេបក្នងតែីសាស្ត្រដែលប្តែល់អនុសាសន៍នេបក្នង
រហាយពារណ៍ចុងក្រោយនេះ ។


Line Segmentation
Script Identification
Font Identification
Word Segmentation
Ground-truth Data
Syllabic Segmentation
Character Segmentation
Feature Extraction
-Pattern matching
-Statistics computation
-Threshold estimation
Classifier selection
Classifier combination
C3
Classifier Combination Wrapper
C1
C2
Training Files
Test Results
Evaluation
Training Phase
Testing Phase

Evaluation *(English Document)*

2972 Characters
202 Errors
93.20% Accuracy

0 Reject Characters
0 Suspect Markers
0 False Marks
0.00% Characters Marked
93.20% Accuracy After Correction

Ins	Subst	Del	Errors	
0	0	0	0	Marked
1	97	104	202	Unmarked
1	97	104	202	Total

Count	Missed	%Right	
439	2	99.54	ASCII Spacing Characters
99	37	62.63	ASCII Special Symbols
83	17	79.52	ASCII Digits
2342	34	98.55	ASCII Lowercase Letters
1	0	100.00	Latin1 Lowercase Letters
8	8	0.00	General Punctuation

2972 98 96.70% Total Character Accuracy

Errors	Marked	Correct-Generated : ALL Confusions
35	0	{}-{ }
28	0	{,}-{ .}
10	0	{,}-{.}
6	0	{<201C>}-{"" }
6	0	{<201D>}-{ ""}
5	0	{m}-{[""p[}
5	0	{oo,o}-{00 . }

Gazetteer and Cambodian OCR

- Multi-lingual
- Multi-column
- Multi-font

اس دستاویز کی مزید کاپیاں آڈیو کیسیٹ پر اور بڑے حروف کی چھپائی میں اور کمیونٹی کی زبانوں میں طلب کیے جانے پر دستیاب ہیں، برائے مہربانی اس پتہ پر رابطہ کریں:

এই ডকুমেন্ট-এর (দলিল) অতিরিক্ত কপি, অডিও এবং বড়ো ছাপার অক্ষর আকারে এবং সম্প্রদায়গুলোর ভাষায় অনুরোধের মাধ্যমে পাওয়া যাবে, অনুগ্রহ করে যোগাযোগ করুন:

Gheibhear lethbhreacan a bharrachd ann an cruth ris an èistear, ann an clò mòr agus ann an cànain coimhearsnachd. Cuir fios gu:

इस दस्तावेज/कागज़ात की और प्रतियाँ, माँगे जाने पर, ऑडियो टैप पर और बड़े अक्षरों में तथा कम्यूनिटी भाषाओं में मिल सकती हैं, कृपया संपर्क करें:

ਇਸ ਦਸਤਾਵੇਜ਼/ਕਾਗ਼ਜ਼ਾਤ ਦੀਆਂ ਹੋਰ ਕਾਪੀਆਂ, ਮੰਗੇ ਜਾਣ 'ਤੇ, ਆੱਡਿਓ ਟੇਪ ਉੱਪਰ ਅਤੇ ਵੱਡੇ ਅੱਖਰਾਂ ਵਿਚ ਅਤੇ ਕੰਮਿਉਨਿਟੀ ਭਾਸ਼ਾਵਾਂ ਦੇ ਵਿਚ ਮਿਲ ਸਕਦੀਆਂ ਹਨ, ਕ੍ਰਿਪਾ ਕਰਕੇ ਸੰਪਰਕ ਕਰੋ:

此文件有更多備份，如果需要，語音版本和大字體版本及少數種族語言版本也可提供，請聯絡：

يمكن أن تطلب النسخ الأخرى من هذا المستند كالتسجيل الصوتي والخط المكبر ونسخ بلغات أخرى، يرجى الإتصال على:

t Bântéay Méan Cheăy

ភូមិ	UTM-X	UTM-Y	ភូមិ	Phum	ស្រុក	Srŏk	ឃុំ	Khŭm	UTM-X	UTM-Y
1002	303400	1490200	ស្រណាល	Srânal						
1003	303800	1489900	ឡ	Lâ						
1004	304400	1490000	តាមែងពក	Ta Mêng Pôk						
1005	305400	1489000	សំបួរ	Sâmbuŏr						
1006	308200	1487800	ដូនឡឹក	Don Lŏĕk						
1007	308600	1487400	ក្បាលក្រាបី	Kbal Krâbei						
1008	309000	1487500	ស្រះឈូក	Srăh Chhuk						
1009	309600	1487400	ស្រែព្រៃ	Srê Prey						
1010	310600	1484800	ចែកអង្គរ	Chêk Ângkâr						
1011	309600	1483500	ថ្មដប់	Thmâ Dâb						
1101	275900	1478400	តាម៉ៅ	Ta Mau						
1102	276000	1478900	អន្សមចេក	Ânsâm Chék						
1103	276200	1479300	ត្នោត	Tnaôt						
1104	276300	1480100	ប្អរ	Buŏr						
1105	275900	1480500	បុស្សឡោក	Bŏss Laôk						
1106	276100	1481400	សៀ	Soeă						
1107	276000	1481600	បឹងតូច	Bŏeng Toch						
1108	276100	1482000	ផ្លូវជីរលើ	Phlov Dâmrei Leu						
							សៀ	Soeă	276400	1481700

Khmer Romanization

English Caps

Digits

Khmer

Multi-lingual

Multi-column

Multi-font

Challenges

- Embed script-identification at word-level (after word-segmentation)
- Font(s) unknown – both phonetic English (Khmer romanization) and Khmer
- Run different recognizers for each word/script
- Concatenate all the results to generate the text document

Approach

- Read the main ***configuration*** file
- Generate ***an instance of recognizer*** for each script present on the multilingual document
- ***Segment*** the document into zone→line→word level
- Apply ***script identification*** at word level
- ***Call specific recognizer object*** for the given script
- ***Recognize*** the word
- ***Merge*** the results of different recognizers into the same page

Linear vs. Non-linear

- Segmentation of document into words is *non-linear (bottom-up approach)*
- Words are organized linearly (in text-flow fashion) using spatial parameters
- Script id and recognizer should be called after word-organization
- Then the instances of recognizers are created
- Once recognized, words should be pasted in linear fashion

Word-segmentation

t Bântéay Méan Cheăy

ល.រ	UTM-X	UTM-Y	ភូមិ	Phum	ស្រុក	Srŏk	ឃុំ	Khŭm	UTM-X	UTM-Y
307	299500	1495200	ខ្លាខាំឆ្កែ	Khla Khăm Chhkē						
308	300200	1492500	បឹងវែង	Bŏĕng Vēng						
101	329800	1524700	រំដួង	Rŏngvéan	ភ្នំស្រុក	Phnum Srŏk	ណាំតៅ	Năm Tau	327300	1527500
102	327200	1527700	ថ្មីខាងត្បូង	Thmei Khang Tbong						
103	327200	1528100	ថ្មីខាងជើង	Thmei Khang Cheung						
104	327300	1528800	គោកយាង	Koŭk Yéang						
105	327600	1528700	គោកចារ	Koŭk Chăs						
106	327300	1529500	ច្រាប	Chrab						
107	327400	1530400	កន្តួត	Kântuŏt						
108	327400	1531100	ណាំតៅ	Năm Tau						
109	327300	1531600	ពង្រ	Pŏngrō						
110	328500	1534900	សំរោង	Sâmraŏng						
111	329600	1536400	ក្នាង	Knăng						
112	330800	1538100	ត្នោងខាងត្បូង	Thnóng Khang Tbong						
113	331100	1538500	ត្នោងខាងជើង	Thnóng Khang Cheung						

Results (script id)

Word-ids	Script-id	Choice 1 *(with confidence)*	Choice 2 *(with confidence)*
0007	0 :	0 (-6.587093)	1 (-73.201634)
0008	0 :	0 (1.373798)	1 (-34.786581)
0009	0 :	0 (-0.507051)	1 (-41.932786)
0010	0 :	0 (-1.324335)	1 (-23.058608)
0011	0 :	0 (-4.364201)	1 (-54.584848)
0012	1 :	1 (-4.044712)	0 (-34.130679)
0013	0 :	0 (-6.615302)	1 (-36.660218)
0014	1 :	1 (-3.540809)	0 (-71.135889)
0015	1 :	1 (1.695183)	0 (-1.565632)
0016	0 :	0 (-5.045144)	1 (-44.335387)
0017	0 :	0 (-0.168188)	1 (-14.357100)
0018	0 :	0 (1.123105)	1 (-10.059393)
0019	0 :	0 (1.713712)	1 (-45.626832)
0020	0 :	0 (-4.052548)	1 (-7.284456)

Ongoing Work

- How to organize the recognized text from different recognizers in multi-column format?
- Evaluation of Khmer script
- Better recognition engines/strategies need to be applied for the degraded documents
- Scanning the documents at a much higher resolution

Testing Data

Training Data

- Bootstrapping the low recognition rates with false script id results
- Using structure of the document to aid script id

Thanks!