The State of the Art in Thai Language Processing

Virach Sornlertlamvanich

Information R&D Division

National Electronics and Computer Technology Center (NECTEC)

THAILAND

virach@nectec.or.th

Introduction to Thai (1): Morphology

Running text (a paragraph):

วิวัฒนาการทางพันธุวิศวกรรมซึ่งเป็นส่วนหนึ่งของเทคโนโลยีชีวภาพ ได้เจริญรุดหน้าไปอย่างรวดเร็ว จนสามารถทำให้เกิดสิ่งมีชีวิตสายพันธุ์ใหม่ ที่เป็นผลมาจากการตัดต่อยีน ซึ่งเราเรียกเจ้าสิ่งมีชีวิตเหล่านั้นว่าสิ่งมีชีวิตแปลงพันธุ์หรือจีเอ็มโอนั่นเอง ปัจจุบัน ความขัดแย้งทางความคิดเกี่ยวกับจีเอ็มโอยังรุนแรงทั่วโลก การสร้างความเข้าใจในเรื่องนี้จึงมีความสำคัญอย่างยิ่ง

- Writing in 4 levels

- No. of characters
  46 consonents; 18 vowels;
  4 tones; 9 symbols; 10 digits
- No word boundary
  Ex:**“GODISNOWHERE”**
  1) God is nowhere
  2) God is now here
  3) God is no where

Introduction to Thai (2): Syntax

- No explicit sentence marker
  - *space character for pausing*
- Sentence pattern
  - (S) (V) (O)

    Ex: ฉัน เห็น เขา

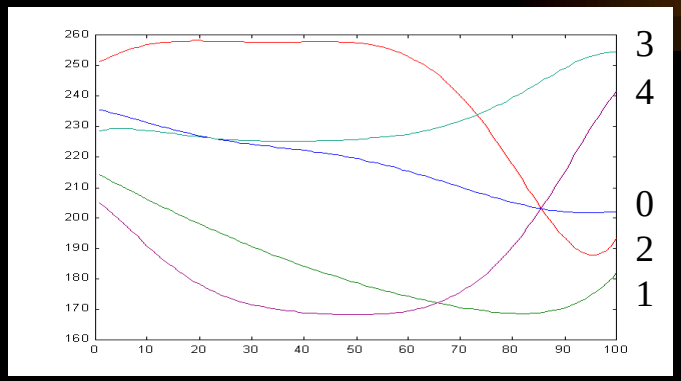
    (I) (saw) (him)
- No inflection forms
  - tenses

    use adverbs and auxiliary verbs
  - plural or singular nouns

    use quantifiers, classifiers or determiners
  - subject-verb agreements
- No syntactic marker
  - *word position*

Introduction to Thai (3): Phonology

- Tone: Thai has 5 tones. Different tones in Thai covey different meanings.
  Ex: สวย (suay4) = beautiful ซวย (suay0) = terrible



Thai Tones

- No liaison:
  A word has the same pronunciation, no matter wherever it is.
- Linking pronunciation:
  ตุ๊กแก (gecko) = tuk4 - kae -> ตุ๊ก = tuk4
  ตุ๊กตา (doll) = tuk4 - ka1 - ta0 -> ตุ๊ก = tuk4 - ka1
  (grapheme to phoneme conversion)

Introduction to Thai (4): Summary

- Simple grammar
  - easy for generation
  - hard for analysis and recognition
- Sharable problems among some Asian languages
  - word segmentation
  - indexing for IR
  - lexical acquisition
  - tone recognition and generation

Research on Thai Language Processing (1)

- Dictionary
  - manually created
  - corpus-based lexical acquisition (COLING2000)
    applying C4.5 on the following language features:
      - left/right mutual information
      - left/right entropy
      - string frequency
      - string length

    yields a result of 85% precision and 56% recall
- Word Segmentation
  - longest matching (92%)
  - maximal matching (93%)
  - part-of-speech n-gram (96%)
  - machine learning (97%)

Research on Thai Language Processing (2)

- Sentence Segmentation
  - part-of-speech trigram (85%)
  - machine learning (89%)

- Information Retrieval
  - current state technology: word based
  - ongoing research: string based, semantic search
    Thai Character Cluster (TCC) based indexing (IRAL2000)

character	เ - ป - อ้ - า - ห - ม - า - ย
cluster (TCC)	เป้า - หมา - ย
word	เป้าหมาย or เป้า - หมาย

Research on Machine Translation (1)

- Difficulties
  - no inflection, grammatical case marker
  - no word and sentence boundary
  - concept unification and decomposition:

    Ex: whiten (English)

    make (something) white (Thai)

- CICC Project (1987-1995)
  - interlingual approach multilingual machine translation
  - 5 languages: Thai, Chinese, Malay, Indonesian and Japanese
  - 50% of average accuracy

Research on Machine Translation (2)

- ParSit (1998 - present)
  - collaboration of NEC (Japan) and NECTEC
  - semi-interlingual approach English-Thai MT
  - June 2000 available to the public, web-based MT
    http://come.to/parsit
  - 80% of accuracy
  - ongoing research: quality improvement, Thai-to-English translation

- UNL (Universal Networking Language) (1996 - Present)
  - United Nations University and 14 countries
  - semantically tagged document to facilitate language independent document processing

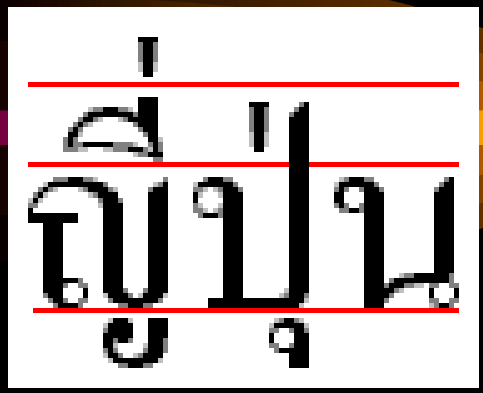
Research on Speech

- Speech Recognition
  - difficulties: *tone recognition*
  - current state technology: *isolated word recognition, speaker identification*
  - ongoing research: *continuous speech recognition*

- Speech Synthesis
  - difficulties: *interplaying between tones and intonation*
  - current state technology: *demisyllable-concatenation based synthesis with tone generation*
  - ongoing research: *smoothing, prosody*

OCR Research

- Difficulties
  - characters in various sizes
  - character-level alignment
  - no word boundary



- Current State Technology
  - neural network

    (95% for registered font; document without images)

- Ongoing Research
  - language model in post-processing
  - documents with text and images

Language Resources

- Text Corpus
  - ORCHID Corpus (1997) supported by CRL Japan
    - 160 documents; 5.75 MB; 311,426 words
    - part-of-speech tagged
    - available for research
  - Difficulties
    - *common understanding in sentence, word and the tags*
- Speech Corpus
  - Large Thai Speech Corpus (2000 -)
  - Collaboration of Advanced Telecommunication Research Institute (Japan), universities and NECTEC
  - to be available for research in 2002
  - grapheme to phoneme conversion