TIDES

***T**ranslingual **I**nformation **D**etection, **E**xtraction, and **S**ummarization*

Why TIDES?

- 200M Web pages as of 7/97
- 1 M terabytes of audio / year
- Uncounted printed matter
- Foreign language information growing faster than English

ITC

National Security

தலைமைச் செயலகம்
தமிழீழ விடுதலைப் புலிகள்
தமிழீழம்
13.05.1998

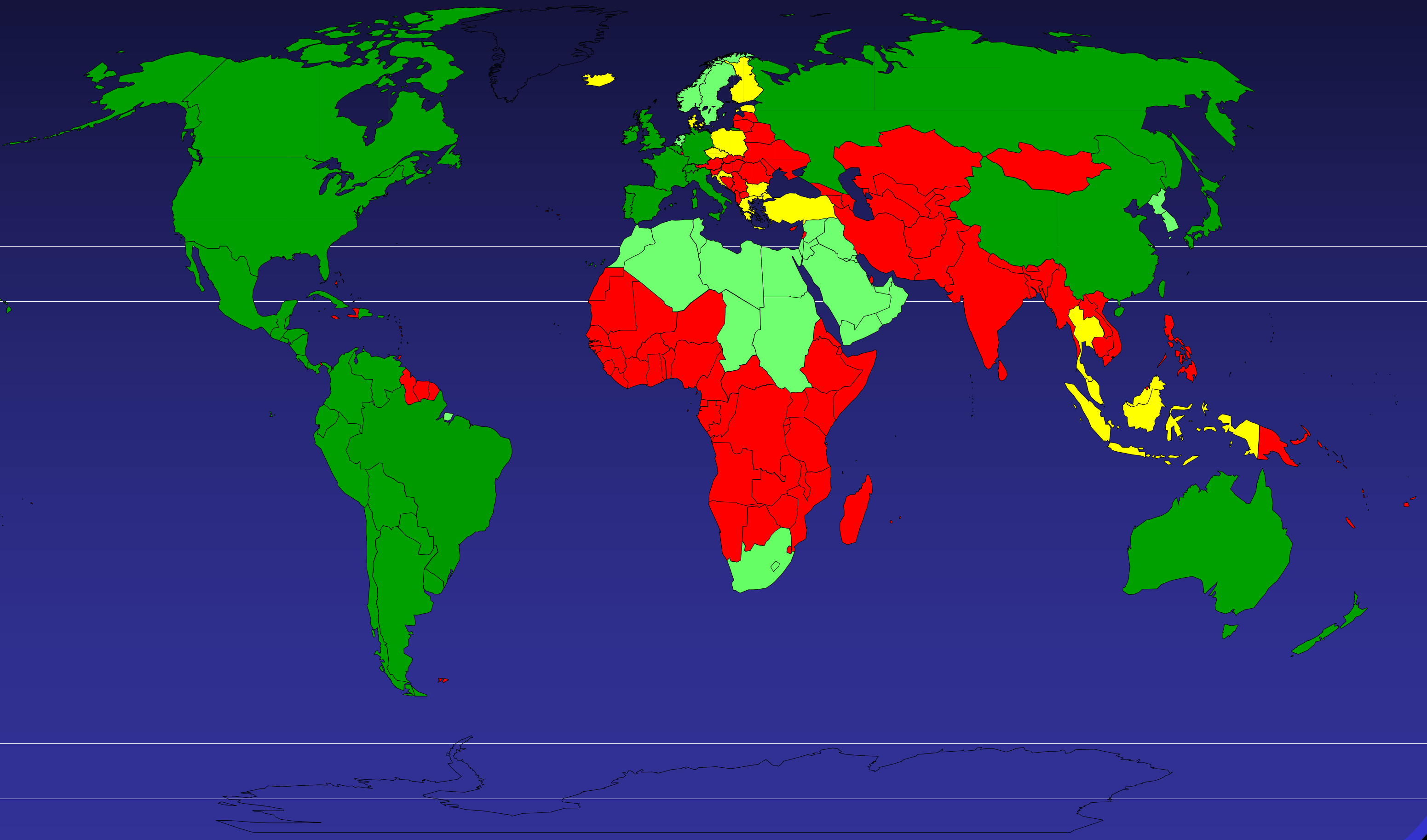
எமது தேசிய விடுதலைப் போராட்ட வரலாற்றில் இ
முக்கியத்துவம் வாய்ந்த நாள். எமது எதிரியான சிறீ
மிகப் பெரிய படையெடுப்பான "ஜெயசிக்குறு" இரா
எதிர்த்து நின்று போராடி, இன்றுடன் ஓராண்டு பூர்
மாத காலத்திற்குள் முடிந்துவிடுமென போர்ப்பறை
பிரச்சார எடுப்புடன் ஆரம்பமான இப்போர் நடவடிக்
வருடமாகியும் இன்னும் முடிவுபெறாது இழுபடுகிறது.
எட்டிவிட்ட ஒரு தனிச்சமர் என்ற ரீதியில், தமிழீழப்
வரலாற்றில் மட்டுமின்றி உலகப் போரியல் வரலாற்
நீண்டதொரு சமராக இது முக்கியத்துவம் பெறுகிறது
படையெடுப்பை மூர்க்கமாக எதிர்த்துப் போராடி, எதி
நகர்வு வேகத்தை தடுத்து நிறுத்தி, எதிரிப்படைகளை வன்னிக்
காட்டிற்குள் முடக்கி வைத்து உலக இராணுவ வரலாற்றில் ஒரு ஒப்பற்ற
சாதனையை எமது விடுதலை இயக்கம் நிலைநாட்டியிக்கிறது.

ITC

TIDES Goal

- Find and Interpret Information Vital to National Security
  - Retrieve unfamiliar languages
  - Translate into English
  - Extract and correlate content

Machine Translation



Bombs & Warnings

Targets

- Translingual access rivaling monolingual access
- Rapid development of MT for new languages
- Multi-document information extraction and correlation

The Problem

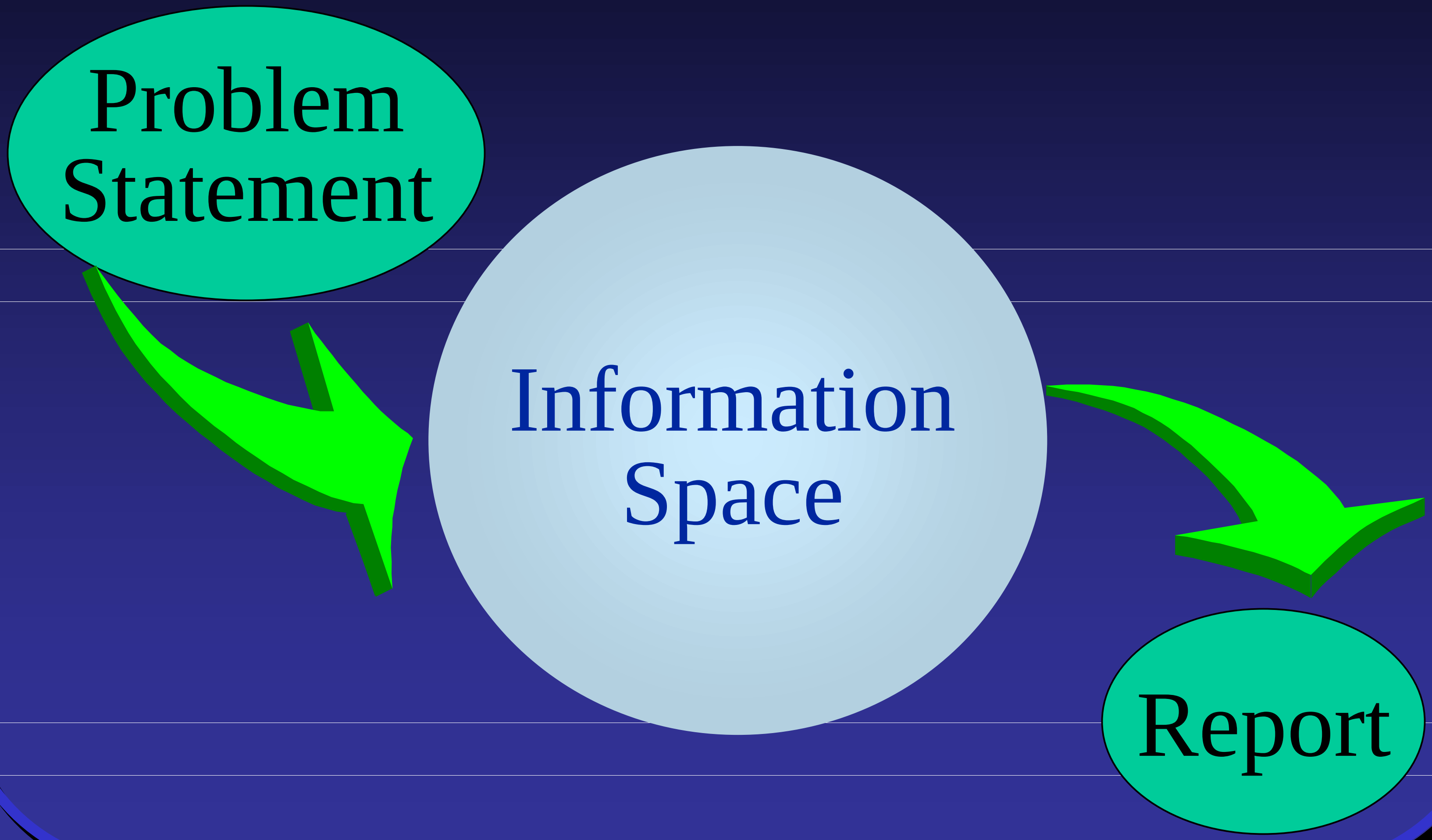
- Exhaustive coverage expected
- Many simmering pots
- Unpredictable flare-ups
- Accurate analysis critical

ITC

The World - 1999

- ~228 Countries
- >6,700 Languages
- >39,000 Language, dialect, and alternate names

Framework

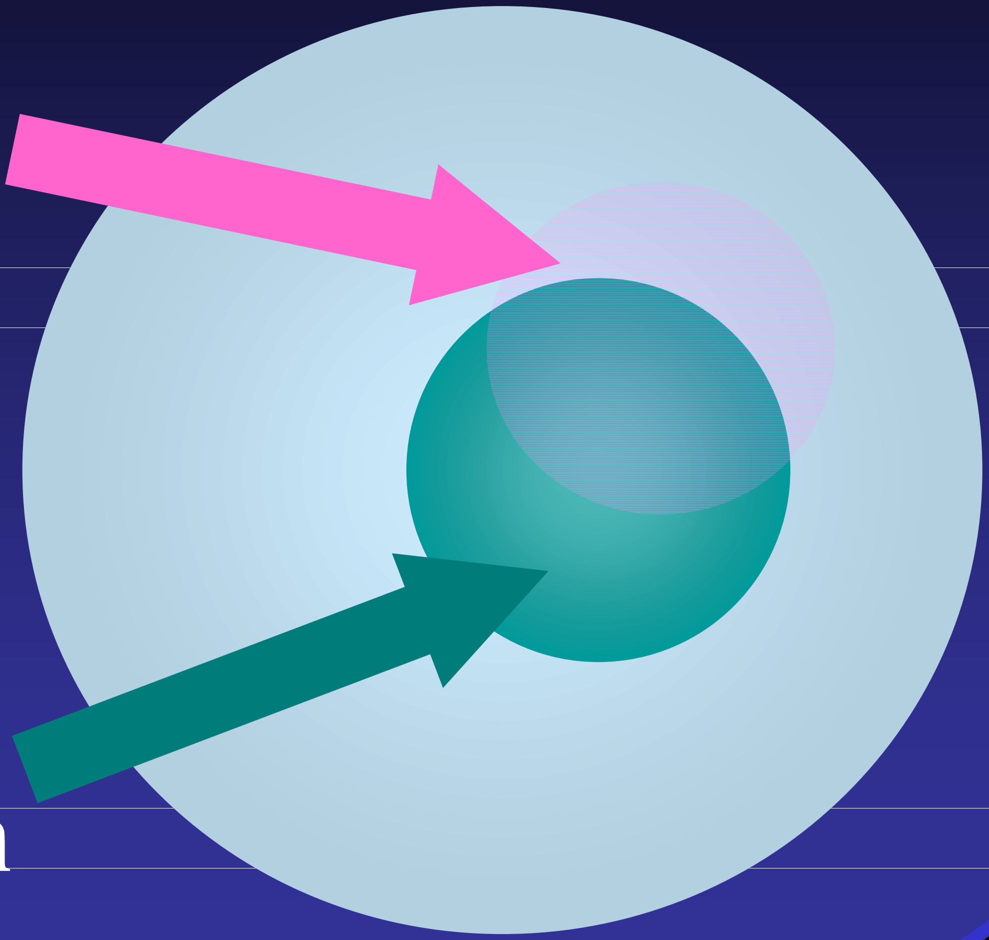


Process Steps

- Information Retrieval
  - Topic Detection
    - Entity Extraction
      - Summary Generation

Information Retrieval

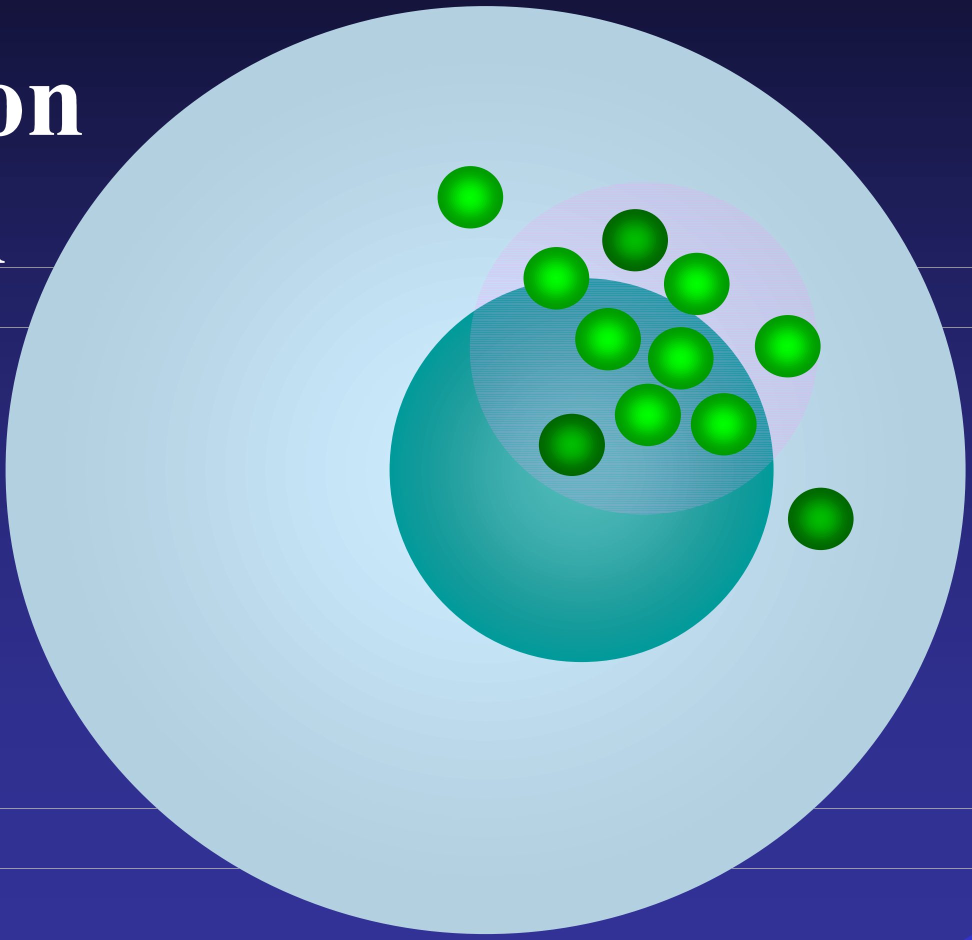
Retrieved Information



Relevant Information

Topic Detection

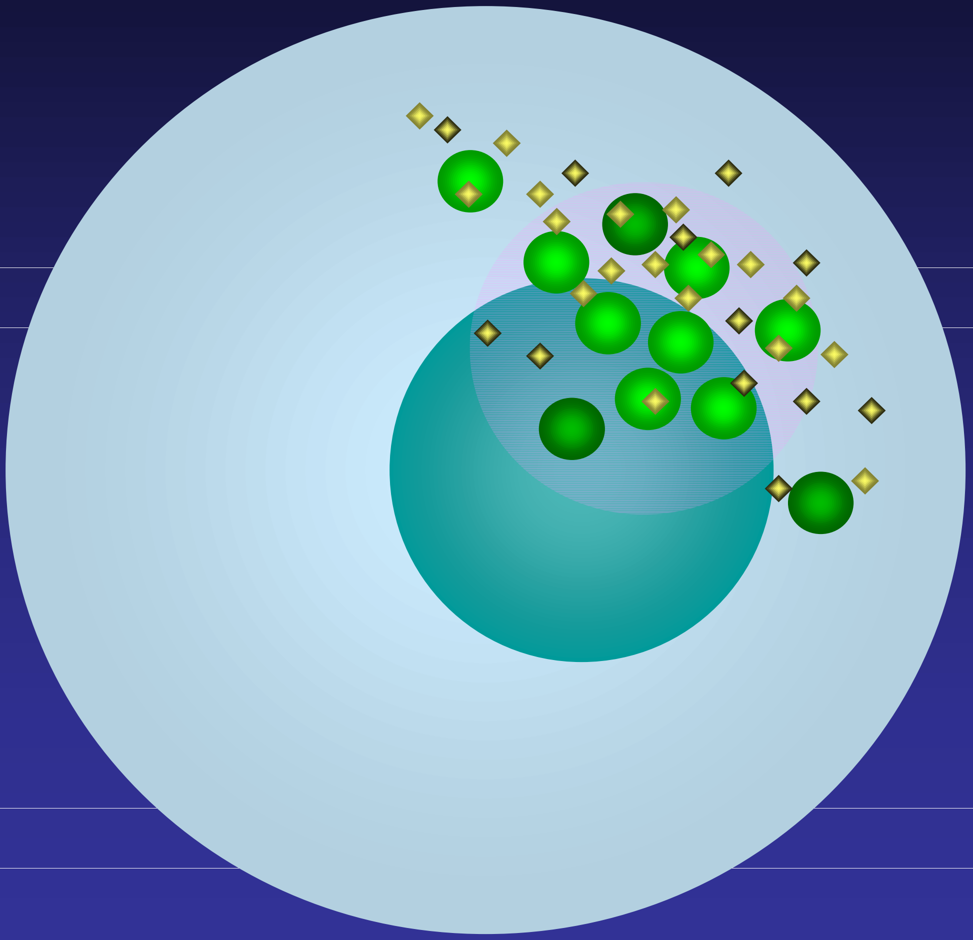
- Segmentation
- Recognition
- Tracking



ITC

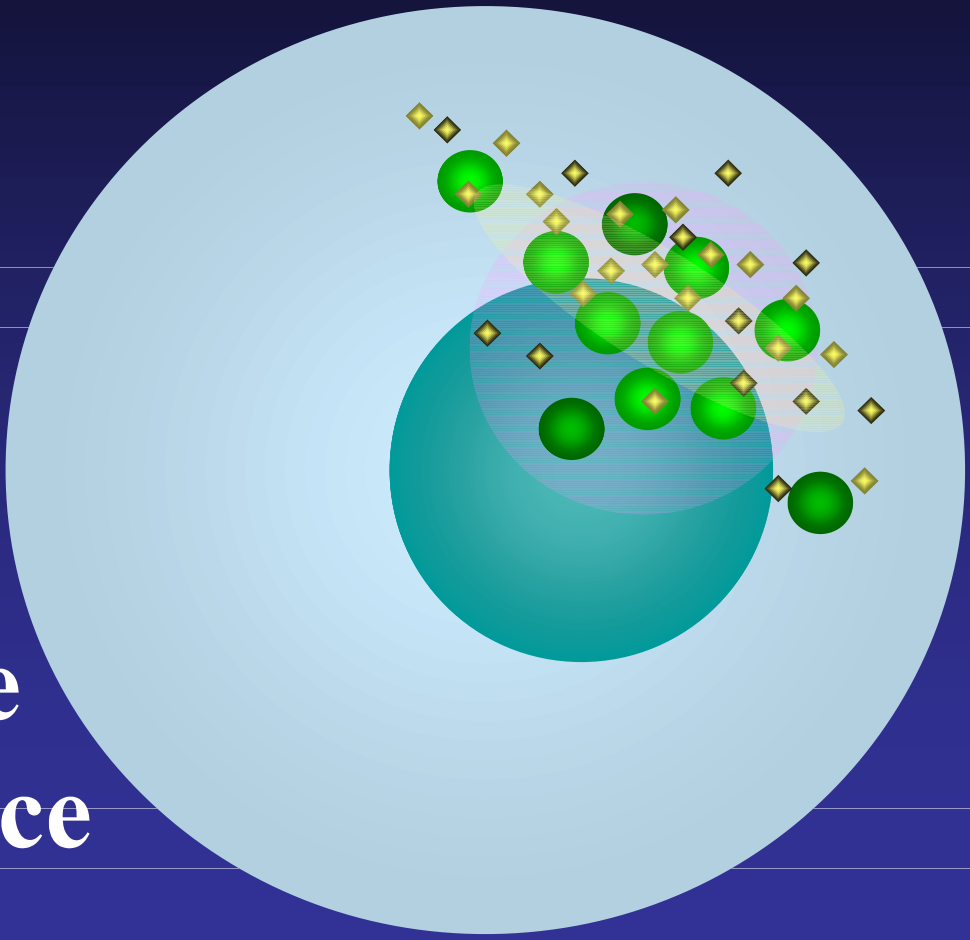
Entity Extraction

- Names
- Places
- Events



Summarization

- Type
- Content
- Perspective
- Performance



Environment

- Large information space
- Human knowledge, patience, and labor
- Relevance feedback
- Monolingual (English)

Beyond English

- Query translation
- Document translation
- 50% performance of monolingual retrieval

Exploiting Feedback

- Relevance feedback
- Topic unification
- Content threading
- Multidocument summarization

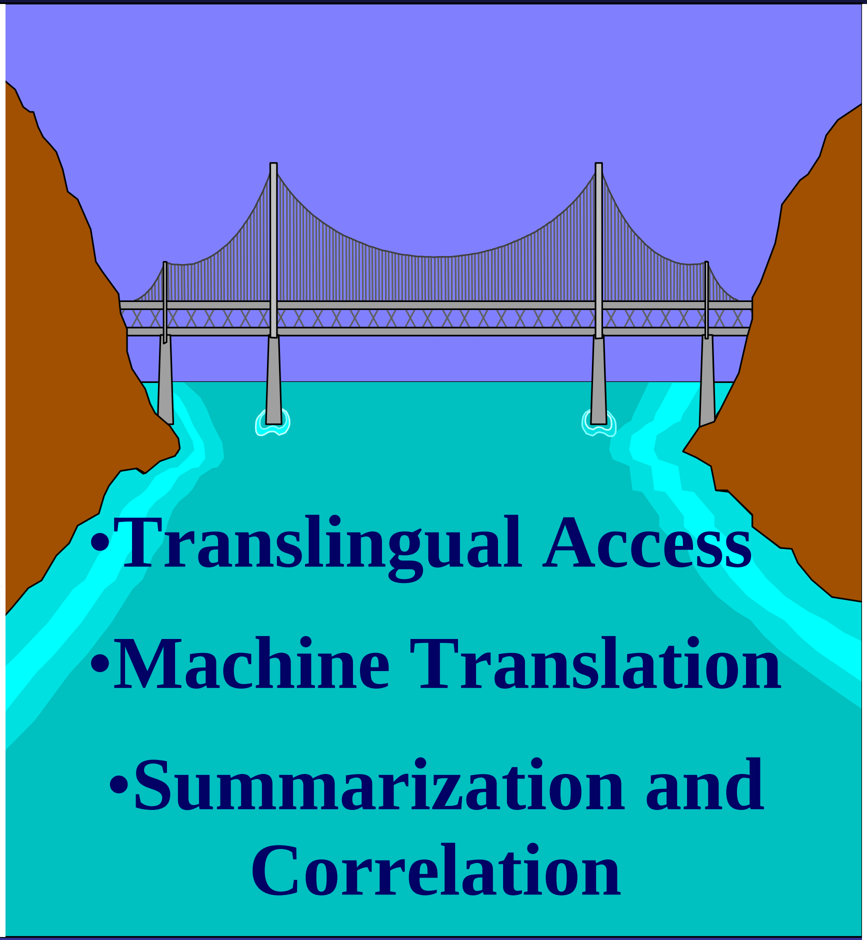
3-Year Goals

- Improved translingual IR
- Rapid shift to new language
- Multilingual topic recognition
- Multidocument summarization

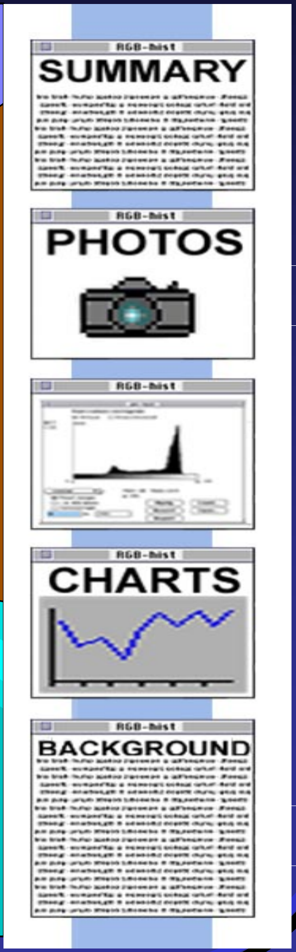
5-Year Goals

- 30+ languages
- Multilingual entity correlation
- Multilingual templates
- Multilingual summarization

TIDES



- Translingual Access
- Machine Translation
- Summarization and Correlation



ITC