Natural language, machine translation, and the democratization of knowledge

Mike Gasser
Collaborators

- Steve Hockema
- Matt Kane
- Amr Sabry, Ahmed Hamed

Related projects
- Information Smuggling
- Author Trustworthiness
What this talk is about

• Why we do research
• Inter-relationships among
  - Knowledge
  - Language
  - (Power)
  - Informatics
What this talk is about

1. Knowledge inequity and the Linguistic Digital Divide
3. Language and the Global Human Knowledge Base
Knowledge: inside and outside

• Knowledge “inside” the person
  - Competence and performance (Chomsky, etc.)
  - Innate and acquired knowledge
  - Knowledge and cognitive representation
  - Tacit knowledge, knowing how (Polanyi, Anderson)

• Knowledge “outside” the person: information
  - Codified knowledge: spoken and written
  - Images, sounds
  - Demonstration
  - Can be shared, passed on, exchanged; permits collaboration

• Inside and outside (Clark)
Codified knowledge
Codified knowledge
Collaborative knowledge
Collaborative knowledge
Collaborative knowledge
Writing
Languages
The distribution of codified verbal knowledge
The Knowledge-Based Society/Economy (David & Foray)

- Acceleration of knowledge production
- Shift to knowledge-intensive activities: R&D, education, software
- Economic disparity not a matter of physical resources but of creating new knowledge and incorporating it in equipment and people
Knowledge inequity

- Knowledge is created, transmitted, and acquired everywhere.
- Some kinds of technical knowledge are essential for particular kinds of progress.
- These kinds of knowledge are disproportionately created in and disseminated to particular regions of the world.
- Easily accessible knowledge is largely western.
Knowledge inequity

• Causes of inequity
  - Geography, historical accident
  - Differential access to education
  - Imperialism (legacy of, continuing)
  - Language

• Effects of inequity (in the global KBS)
  - Continued (or exacerbated) economic disparities based on geographical region, economic development, ethnicity, gender, class background
  - Failure to address pressing national and global issues that require collaboration
Democratization of Knowledge as an international goal

  - 3: “the ability for all to access and contribute information, ideas and knowledge”
  - 8: “stimulate respect for cultural identity, cultural and linguistic diversity, traditions and religions and foster dialogue among cultures and civilizations”

- The potential for the Internet to address DoK
Language and participation in the KBS

- **Codification**
  - Literacy
  - Access to technology

- **Collaborative creation**
  - Shared language
  - Interpreter
  - Access to technology

- **Transmission and acquisition**
  - Shared language
  - Interpreter
  - Access (technology and information)
Language and the Democratization of Knowledge

- **Literacy**
  - ~82%
  - Definition, distribution, implications

- **Access**
  - (Global) Digital Divide
  - Intellectual property rights and the Open Access movement

- **Shared languages or interpreter**
The world’s languages

- 6-7,000 languages are spoken in the world \((Ethnologue)\); about 1/3 are written.

- ~400 languages are spoken as native languages by 1,000,000 or more people, 90% of the world’s population.

- Many people are fluent speakers of second languages, so perhaps 100 are spoken by 90%. 
• One language dominates (~70% of web pages). 12 languages account for ~97% of all web pages. (O’Neill, Lavoie, Bennett)
Language and the Internet (Paolillo)

- Even some communities that share a language other than English (e.g., Panjabi speakers) use English for email and chat.
- Even for languages such as Spanish, resources may focus on music, dancing, food, shopping (in fact may be catering to “cultural tourists”). (Clark & Gorski)
- The Linguistic Digital Divide: separates speakers of privileged languages from others and privileged speakers from others within their linguistic communities.
The Linguistic Digital Divide

- Lack of documents
- Lack of computational resources
- Linguistic bias in “non-linguistic” resources
- Lack of power and financial resources
- Linguistic imperialism, chauvinism
- Lack of users
Bridging the LDD

- Have everybody learn English.
  - It doesn’t work. (Brock-Utne: *The Recolonization of the African Mind*)
  - It relegates all other languages to a secondary role: violates WSIS Principle 8.

- Create documents in (and tools for) under-represented languages.
Translation and the LDD
Translation and the LDD
Registers within a language

- Formal register
  *In addition to a slower rate of new site creation, the rate at which existing sites disappear may have increased.*

- Informal register
  *So people are creating new sites more slowly, and sites already on the Web may also be going away more quickly.*
The L$^3$ Project

- 100 languages, 9900 translation pairs
- Few or non-existent computational resources for most languages
- Statistical MT that learns the grammar and lexicon of languages as it learns to translate between them and that can also make use of explicit linguistic generalizations
- Sharing of knowledge across translation pairs
- Incremental training, beginning with simple language
- Initially rudimentary translations that improve with more training and feedback from users
The basic idea

- Start with pairs of sentences: sequences of morphemes.
- Meaning is distributed. Sentences correspond to sentences.
- On the basis of co-occurrence of units within and between languages, with some probability create links between them, which become units in their own right.
- Strengthen units when they recur; eliminate them if they fail to recur after awhile.
- At processing time, given some units, access others on the basis of the conditional probabilities derivable from the units’ strengths.
Related work

- Edelman et al. (2002-5): Chorus of Phrases, ADIOS
- Dennis (2001-5): Syntagmatic-Paradigmatic Model
- Dependency grammars
  - Tesnière, Hudson, Mel’čuk
- Statistical MT
Morphological parsing

- saw $\rightarrow$ see + PAST
- vimos ‘we saw’ $\rightarrow$ ve- + -mos + PAST
- xoijxilo ‘we saw you’ $\rightarrow$ x- + oj- + ix- + -il- + -o
- እለየናችሁም ‘we didn’t see you’ $\rightarrow$ እል + እየ + PAST + ከ + ከትሱ + የም
- Hand-coded parsers and learned parsers
- Generation
Learning the properties of one language: syntagmatic links

he put on a red shirt
Learning the properties of one language: syntagmatic links
Learning relationships between languages: mapping links

English: he put on a red shirt

Español: se puso una camisa roja
Learning relationships between registers

Formal English:
- in
- the
- unlikely
- event
- of

Informal English:
- in
- case
- there's
Translation

- Given a source sentence $S$, find the target sentence $T$ that is most likely.
- Generate candidate $T$s using mapping links from elements in $S$.
- For each $T$, estimate its goodness using Bayes’ Law (Brown, Della Pietra, Della Pietra, Mercer)

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)}$$

$$\hat{T} = \text{argmax}_T[P(S|T)P(T)]$$

[to program]
Translation

- A (possibly very bad) translation is placed on a Wiki, together with the original.
- Users correct the translation, providing feedback to L^3.
- A fake example.
Learning categories

he put on a coat

he put on a shirt

jacket
Learning shared “concepts”
Learning shared “concepts”

potato

papa

じゃがいも
Learning shared “concepts”
Learned shared “concepts”
Learned shared “concepts”
Implementation

- Contact potential users within the language community to assess interest and needs.
- Hunt for existing data (dictionaries, monolingual texts, bilingual texts) and gather more if necessary.
- Implement morphological parser and generator to the extent possible.
- Train — get feedback, correct — train ...
- First projects
  - Mayan languages in Guatemala, Asociación Ajb’atz’ Enlace Quiché
  - Ethiopia: Amharic, Oromo, Tigrinya
Some qualms about MT

• It doesn’t work.
  - Narrow domains
  - Training
    • Incremental
    • Improves with amount of data, storage space, feedback from users
  - Theory
  - An empirical question

• It presupposes solutions to encoding problems.
• It’s labor intensive.
• It requires large amounts of data.
Towards the Global Human Knowledge Base

- Archiving human knowledge
- Non-verbal knowledge: images, sounds
- Verbal knowledge
  - How should it be represented?
Knowledge and language: one view

• Knowledge (at least some of it) is represented in a universal Language of Thought: “conceptual representations”.

• Language acts as (just) a kind of interface between conceptual representations and other people (Fodor, Chomsky, Pinker, Jackendoff)

• Representing all concepts, facts, conjectures that people can think in terms of a hierarchy of concepts, an ontology.

• Universal ontologies in AI and cognitive science.
Archiving human knowledge: the CYC Project (Lenat)

- “To break the ‘software brittleness bottleneck’ once and for all by constructing a foundation of basic ‘common sense’ knowledge—a semantic substratum of terms, rules, and relations—that will enable a variety of knowledge-intensive products and services”

- Hundreds of thousands of hand-entered assertions

- Uses a knowledge representation language based on FOPC
CYC and the GHKB
Knowledge and language: another view

- Each language is a kind of window on reality, a way of slicing up the world; there is no purely non-linguistic way of representing (propositional) knowledge. (Bowerman, Lakoff, Levinson, Lucy)

- The Sapir-Whorf Hypothesis (linguistic relativism/determinism)
  - Systematic properties of a language have effects of perception, attention, memory
  - New evidence in its favor
Archiving human knowledge: **Wikimedia** projects

- **Wikipedia**
  - “the communal encyclopedia that anyone can edit”

- **Wiktionary**
  - “a collaborative project to produce a free, multilingual dictionary in every language,” “the lexical companion to ... Wikipedia”

- **Wikinews**
  - “the free-content news source you can write!”

- **Wikispecies**
  - “an open, free directory of species”
Knowledge as language

- Translation is only possible to different degrees.
- To the extent that there are good correspondences between groups of languages, we have shared "concepts".
- The "hidden layer" of $L^3$ represents a kind of "ontology."
$L^3$ and the GHKB
How do languages differ?

• **Individual words**
  - መጋገር ‘to be too polite, pretending to have already had enough when offered more of something by a host’
  - ወደ ‘to stay behind’ (sort of)
    - ወደ I stayed behind. I didn’t make it.
    - ወደ I’ll stay behind without eating. = I probably won’t eat.
    - ወደ I won’t stay behind without eating. = I’ll probably eat.
    - ወደ I remained stay without eating. = I didn’t get around to eating.
How do languages differ?

• **Representation of a whole domain**
  - Space in Tzeltal
    - Geocentric vs. egocentric perspective
      - *The bottle is “north” of the cup.*
    - Allocentric body-part terms for contact relations
      - *The fly is on the nose/ear/mouth/bottom of the jug.*
    - Very specific positional expressions
      - *The apple is “bowl-sitting” at the bowl.*
      - *The arrow crosses acrosswards at the apple.*

• **Grammar**
Conclusions

• Language is power
  - The availability of more documents in under-represented languages and the capacity to translate readily into and out of those languages could help to bring more people into local and global decision making.

• Knowledge is (to some extent) language
  - A statistical analysis of many languages may help give us a clearer picture of what languages share and do not share and contribute to the Global Human Knowledge Base.
Conclusions

• Research and the Real World
  - In artificial intelligence and cognitive science, we are concerned with what it means to "know," to "say," to "mean," to "understand."
  - How could insights from these fields make a difference and what we all "know" (and do) given long-term social-political goals?
Gracias.
Maltyox.
አመሰግናለሁ።