Artificial Intelligence for Social Good

Robert Munro
Idibon
CS 124 / Ling 180:
From Languages to Information
Stanford
January 2016
Idibon: language technology for all languages

25 people: HQ in San Francisco
Diverse: 25+ languages; equal gender ratio including engineering
Advisors: Dan Jurafsky, Chris Manning, Chris Potts, Lukas Biewald, Tim Converse, Louis Monier

Experience:

Joining today: Aneesh and Martha
About me: technology & global development

CEO, Idibon
Global text analytics in 50+ languages
Working with leaders in industry & social good

Industry: CTO / CIO
Energy infrastructure in Liberia and Sierra Leone
Global epidemic tracking
Crowdsourcing and natural language processing for disaster response

Other
Ph.D. in NLP from Stanford
Bicycled 20+ countries
Unstructured text is the world’s greatest source of digital information

“…most of it is text.”

Source: IDC
You could not understand the majority of the world’s data

Source: Internet world statistics
… first languages are even more common in direct messaging, relative to www

Source: Data Center Knowledge, Gigaom, Pingdom
... and more common still in phone-based communications

Source: Data Center Knowledge, Gigaom, Pingdom
Every human communication this year

Source: Ethnologue, Nationalencyklopedin
7% of our communications are digital, most is still direct spoken language
If every online picture is worth a thousand words, it would double social media.
Every 3 months, the world's text messages exceed the word count of every book.

Source: Google Books
Print communication is smaller than anything shown.
Print communication is smaller than anything shown.
Ditto any one social network.
The Twitter “firehose” is about the size of the dot above the “i” in English.

Beyond the processing capacity of most organizations.

Might not be a representative sample of all human activity for your area of interest.
There are more than 6,000 other languages. Only the top 1% are shown.
No language from the Americas made the cut.
Email spam would be larger than every block except spoken Mandarin (官话).

Source: Mashable
Short messages (SMS and IM) make up 2% of the world’s communications.

The largest and most linguistically diverse form of written communication that has ever existed.

# PhDs focused on processing large volumes of short messages in low resource languages?
If the Facebook “like” is a one-word language it is in the top 5% of languages by word count.
Your browser probably won't show Sundanese script (ᮀᮁᮂᮃᮄᮅᮆᮇᮈᮉᮊᮋᮌᮍᮎᮏᮐᮑᮒᮓᮔᮕᮖᮗᮘᮙᮚᮛᮜᮝᮞᮟᮩᮡᮢᮣᮤᮥᮦᮧᮨᮩ᮪᮫ᮬᮭᮮᮯ᮰᮱᮲᮳᮴᮵᮶᮷᮸᮹ᮺᮻᮼᮽᮾᮿᮀᮁᮂᮃᮄᮅᮆᮇᮈᮉᮊᮋᮌᮍᮎᮏᮐᮑᮒᮓᮔᮕᮖᮗᮘᮙᮚᮛᮜᮝᮞᮟanzi)
Sundanese speakers outnumber the populations New York, London, Tokyo and Moscow.

Combined.
You misread “Sundanese” as "Sudanese" which is a variety of Arabic.

We have a blind spot for knowing about the existence of languages.
This is the breakdown of languages that most of our data is moving towards
We expect machines to listen

~5 in every 10,000 words are directed at machines, not people

- Mainly search engines
- Increasingly speech recognition

Credit: 2001, A Space Odyssey
What does it mean to connect?

• Putting smart technology in the hands of everyone on the planet was the easy part.
• Understanding everyone is much more complicated.
• What is our obligation to help those that we can connect with?
An earthquake struck Haiti on January 12, 2010. Most local services failed, but most cell-towers remained functional.
Messages start streaming in

- Fanmi mwen nan Kafou, 24 Cote Plage, 41A bezwen manje ak dlo
- Moun kwense nan Sakre Kè nan Pòtoprens
- Ti ekipman Lopital General genyen yo paka minm fè 24 è
- Fanm gen tranche pou fè yon pitit nan Delmas 31
Messages start streaming in

- Fanmi mwen nan Kafou, 24 Côte Plage, 41A bezwen manje ak dlo
- Moun kwense nan Sakre Kè nan Pòtoprens
- Ti ekipman Lopital General genyen yo paka minm fè 24 è
- Fanm gen tranche pou fè yon pitit nan Delmas 31
- My family in Carrefour, 24 Côte Plage, 41A needs food and water
- People trapped in Sacred Heart Church, PauP
- General Hospital has less than 24 hrs. supplies
- Undergoing children delivery Delmas 31
Mission 4636

“Fanm gen tranche pou fè yon pitit nan Delmas 31”
Undergoing children delivery
Delmas 31
18.495746829274168, 72.31849193572998
Emergency

Message translated, categorized & geolocated

Location is refined & actionable items are identified

“Fanm gen tranche pou fè yon pitit nan Delmas 31”
Undergoing children delivery
Delmas 31
18.495746829274168, 72.31849193572998
Emergency
Global collaboration

2,000 volunteers, transferred to paid workers in Haiti
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”
Local knowledge

Workers collaborating to find locations:

*Dalila:* I need Thomassin Apo please

*Apo:* Kenscoff Route: Lat: 18.495746829274168, Long:-72.31849193572998

*Apo:* This Area after Petion-Ville and Pelerin 5 is not on Google Map. We have no streets name

Feedback from responders:

"just got emergency SMS, child delivery, USCG are acting, and, the GPS coordinates of the location we got from someone of your team were 100% accurate!"

The ability for someone to make a real-time difference at any other place in the world:

*Apo:* I know this place like my pocket

*Dalila:* thank God u was here

‘here’ = anywhere
80,000 messages

FOOD URGENTLY NEEDED AT AVENUE MONSEIGNEUR GUILLOUX
Mesanmi moun monsenye giyou prolonge antre ru snelak dekay?
t ap mouri grangou
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ the people of monseigneur Giyou between sleslak street are dying of hunger. TN monseigneur Giyou is a residential neighborhood near the st Therese school in downtown ~~~~~~~~~~~~~~~~~~~~
80,000 messages
80,000 messages
80,000 messages
80,000 messages
80,000 messages
Local Impact

• US Marines
  – “Saving lives every day.”

• World Food Program
  – “We delivered food to an informal camp of 2500 people that you identified for us.”

• Workers we employed in Haiti:
  – Have collectively earned more than $1,000,000 to date as professional microtaskers since we first trained them
Can we automate part of this process with NLP?

My PhD focused on this problem, and Idibon has helped dozens of disaster response organizations utilize NLP and crowdsourcing.

… but for this presentation, a more global example
Scaling beyond purely manual processing.

Disease outbreaks are the world’s single greatest killer.
No organization is tracking them all.
Human Diseases eradicated in the last 75 years:

Increasing in air travel in the last 75 years:
90% of ecological diversity

90% of linguistic diversity
Reported locally before identification

Simply *finding* these early reports can help prevent epidemics.

H1N1 (Swine Flu) months
(10% of world infected)

H1N5 (Bird Flu) weeks
 (>50% fatal)

HIV decades
(35 million infected)
в предстоящий осенне-зимний период в Украине ожидается две эпидемии гриппа

مزيج من انفلونزا الطيور في مصر

香港现1例H5N1禽流感病例曾游上海南京等地
E Coli in Germany

A package of bean sprouts from the Bienenbuttel farm in Lower Saxony is positive for E. coli. (Buenos Aires Herald)
UNICEF utilizes Idibon to process millions of SMS in 12 African languages.

- **Intent of sender**: (i.e. report a problem, ask a question or make a suggestion)
- **Categorization**: (i.e. orphans and vulnerable children, violence against children, health, nutrition)
- **Language detection**: (i.e. English, Acholi, Karamojong, Luganda, Nkole, Swahili, Lango)
- **Location**: (i.e. village names)
In Summary

We will never again be so under-resourced to process the world’s information.

Natural Language Processing allows us to understand communication at scale.

Questions?