The world inside words: information extraction and labeling in low resource languages through subword models

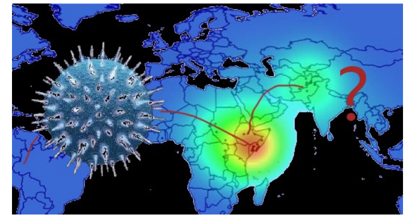
Robert Munro

CEO, Idibon

(research while at Stanford)

Microsoft Research, October 2012

About me



- CEO of Idibon
  - Language technology startup
- Former CTO of epidemicIQ
  - Tracking outbreaks with NLP and crowdsourcing

- PhD from Stanford
  - Computational linguistics
- Coordinator of Mission 4636
  - Emergency response in Haiti
- Power Infrastructure in West Africa
  - Energy for Opportunity / UN
- Traveler
  - 20 countries by bicycle

Acknowledgments

- Chris Manning
- Dan Jurafsky
- Tapan Parikh

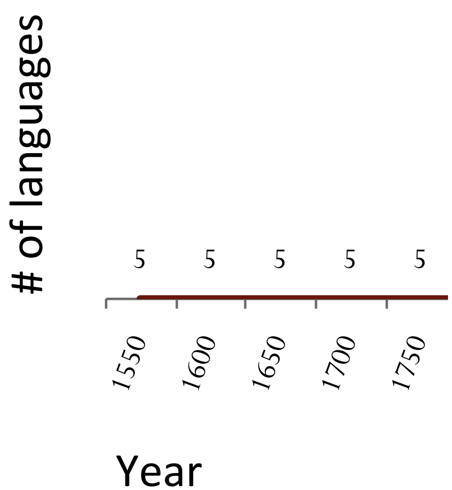
- Stanford NLP
- Stanford Linguistics

Technology for low resource languages

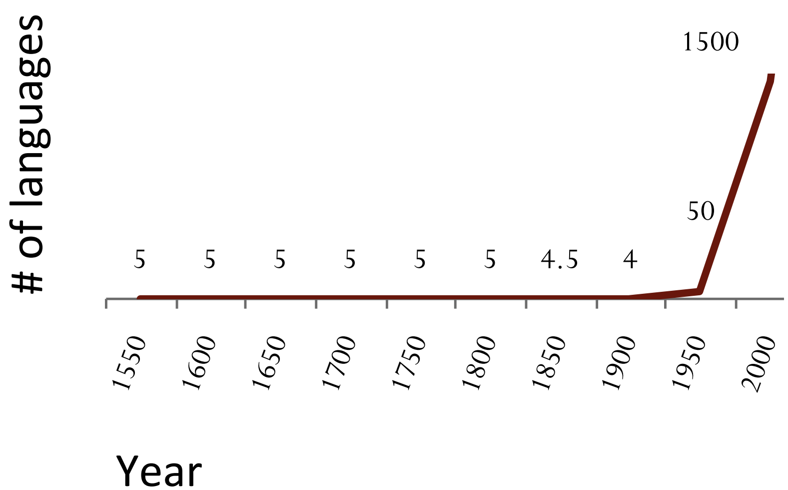
- *Microsoft Translator Hub*
- One of the most important recent advances!
- c/o Will Lewis, Kristin Tolle (MSR Redmond)

- I am interested to hear more about MSR's work using language technologies to augmented textbooks!

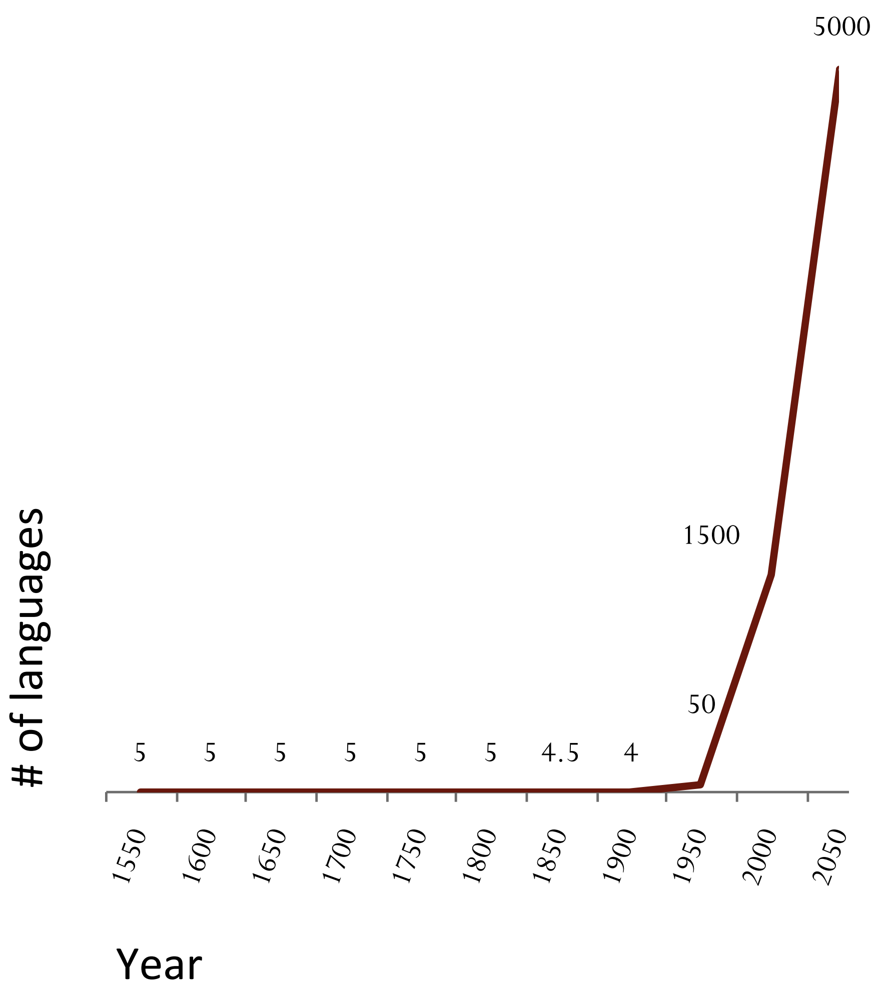
Daily potential language exposure



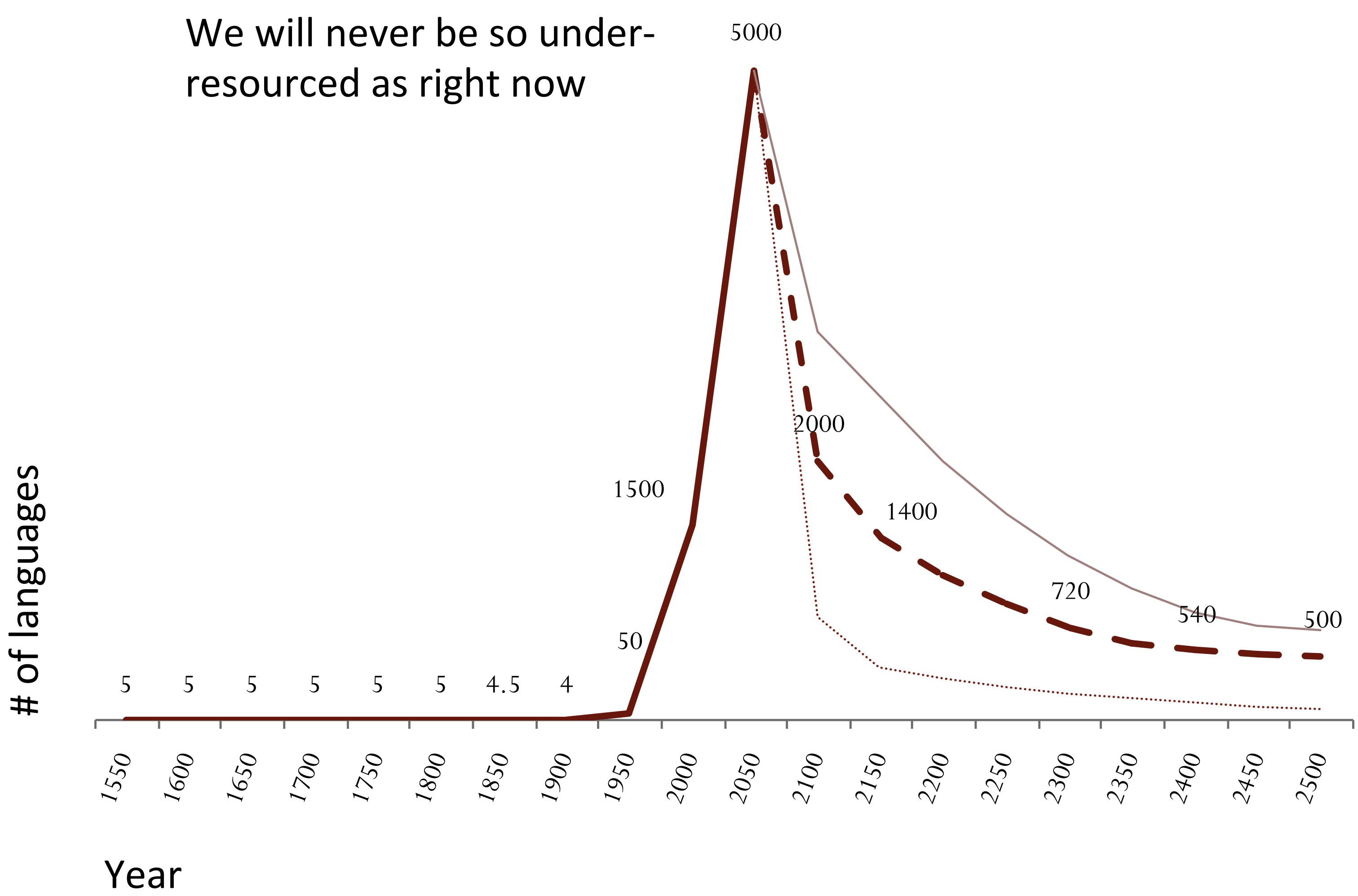
Daily potential language exposure



Daily potential language exposure



Daily potential language exposure



Motivation

- Text messaging
  - Most popular form of remote communication in much of the world [1]
  - Especially in areas of linguistic diversity
  - Little research

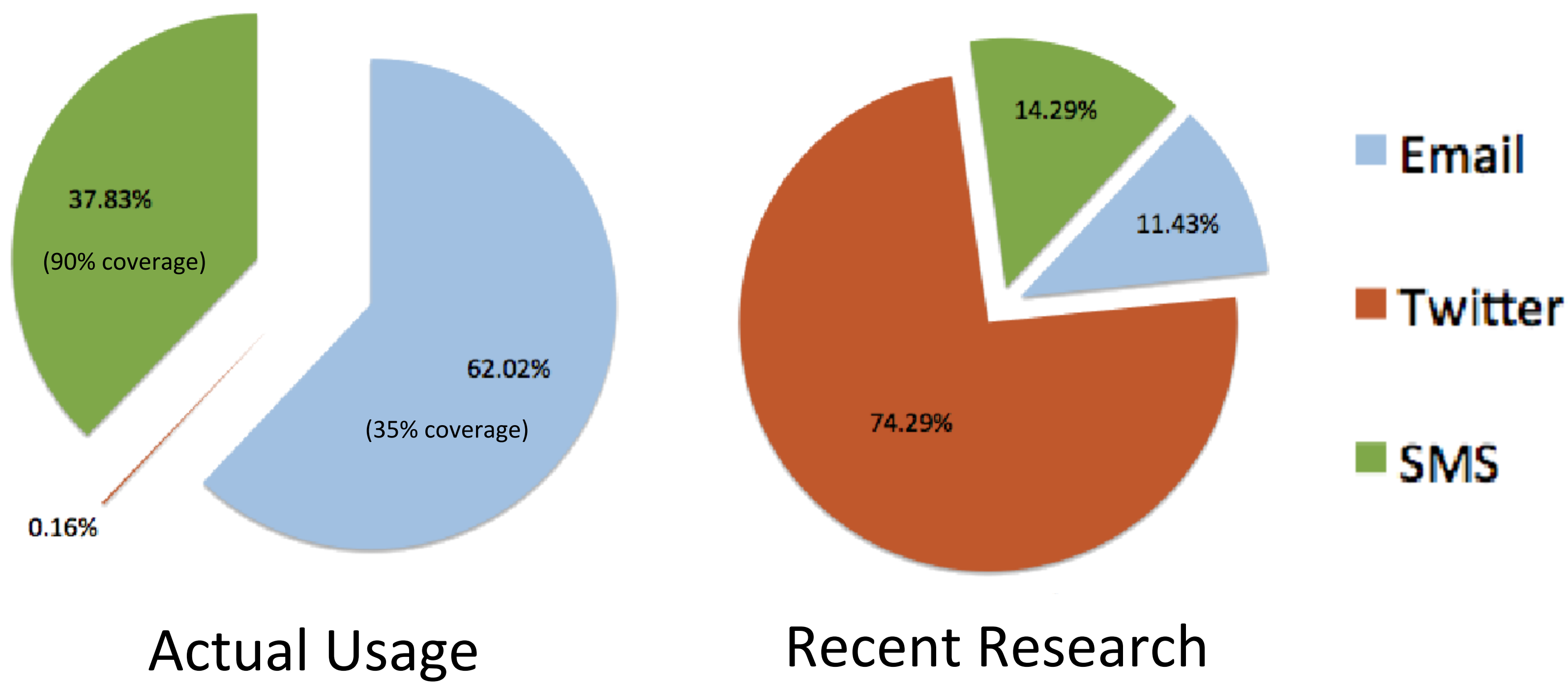
2012 (estimate): 9 Trillion

2007: 5 Trillion

2000: 1 Trillion

[1] *International Telecommunication Union* (ITU), 2012. http://www.itu.int/ITU-D/ict/statistics/

ACM, IEEE and ACL publications



Outline

- What do short message communications look like in most languages?
- How can we model the inherent variation?
- Can we create accurate classification systems despite the variation?
- Can we leverage loosely aligned translations for information extraction?

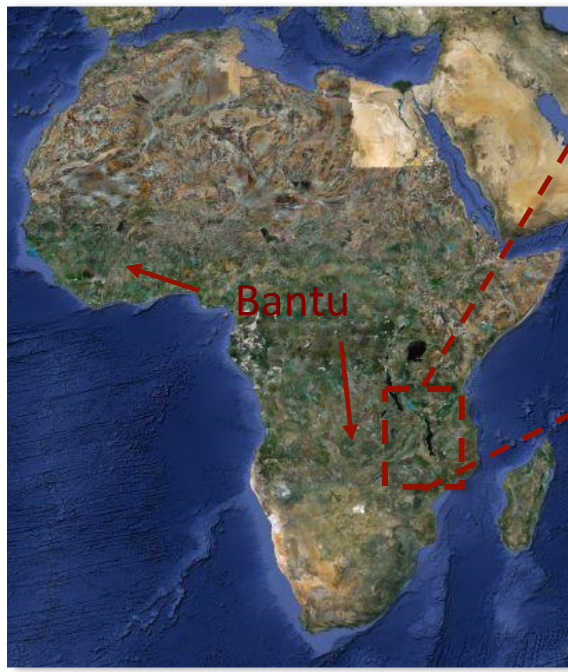
Data – short messages used here

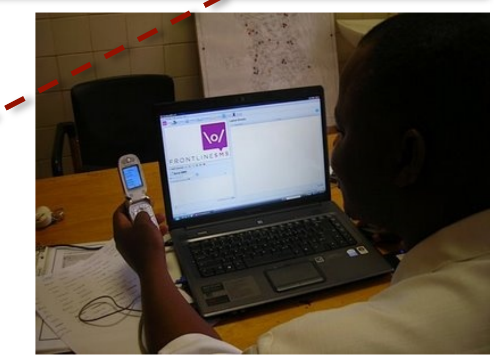
- 600 text messages sent between health workers in Malawi, in Chichewa
- 40,000 text messages sent from the Haitian population, in Haitian Kreyol
- 500 text messages sent from the Pakistani population, in Urdu
- Twitter messages from Haiti and Pakistan
- English translations

Chichewa, Malawi

- 600 text messages sent between health workers, with translations and 0-9 labels

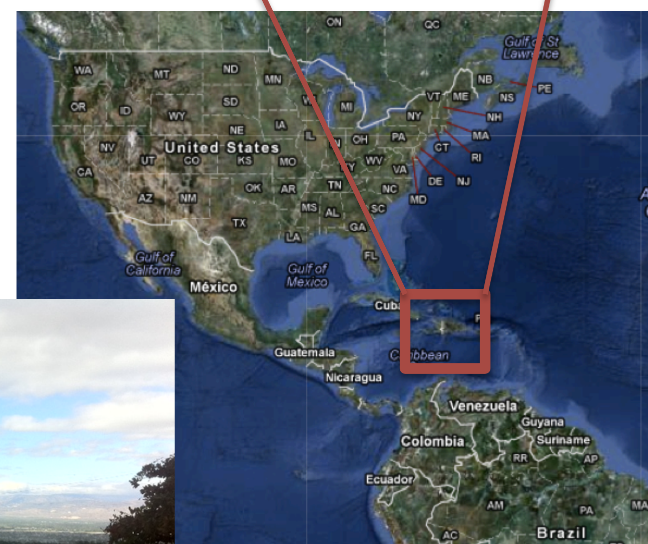
1. Patient-related
2. Clinic-admin
3. Technological
4. Response
5. Request for doctor
6. Medical advice
7. TB: tuberculosis
8. HIV
9. Death of patient

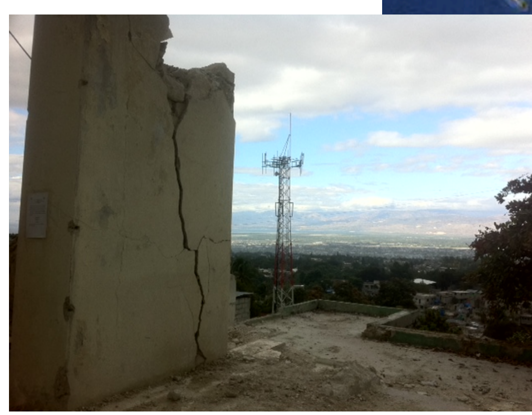




Haitian Kreyol

- 40,000 text messages sent from the Haitian population to international relief efforts (Mission 4636)
  - ~40 labels (request for food, emergency, logistics, etc)
  - Translations
  - Named-entities
- 60,000 tweets





Urdu, Pakistan

- 500 text messages sent from the Pakistan population to international relief efforts
  - ~40 labels
  - Translations
- 1,000 tweets

Outline

- What do short message communications look like in most languages?
- How can we model the inherent variation?
- Can we create accurate classification systems despite the variation?
- Can we leverage loosely aligned translations for information extraction?

Most NLP research to date assumes the standardization found in written English

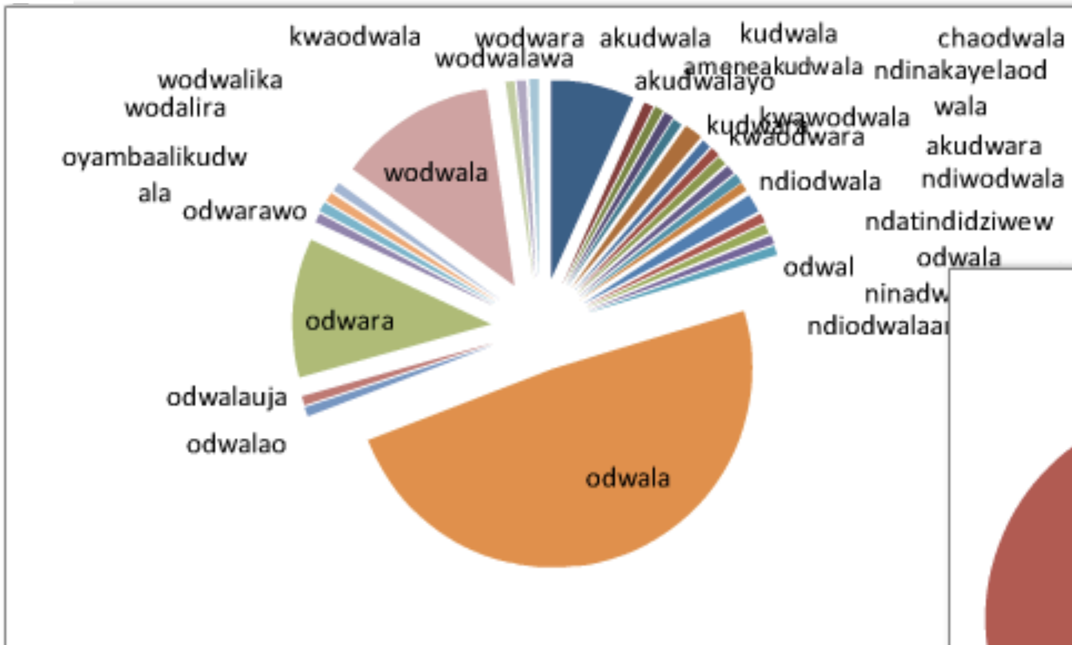
English

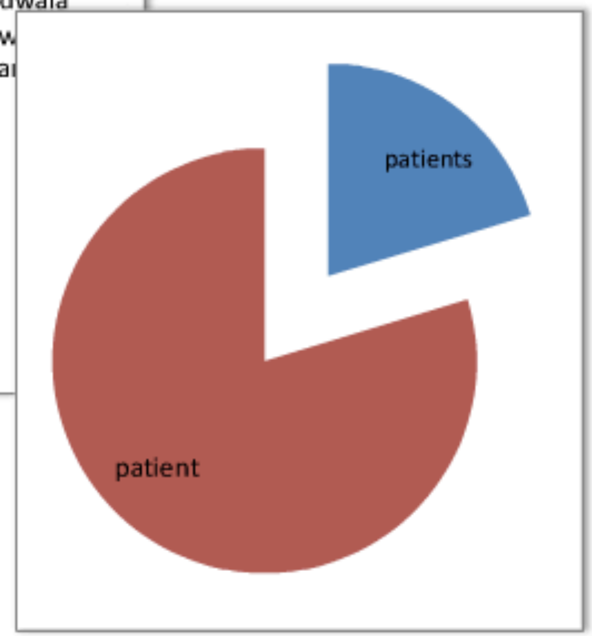
- Generations of standardization in spelling and simple morphology
  - Whole words suitable as features for NLP systems
- Most other languages
  - Relatively complex morphology
  - Less (observed) standardized spellings
  - More dialectal variation
- '*Subword variation*' used to refer to any difference in forms resulting from the above

The extent of the subword variation

- >30 spellings of *odwala* ('patient') in Chichewa
- >50% variants of 'odwala' occur only once in the data used here:
  - Affixes and incorporation
    - 'kwaodwala' -> 'kwa + odwala'
    - 'ndiodwala' -> 'ndi odwala' (official 'ngodwala' not present)
  - Phonological/Orthographic
    - 'odwara' -> 'odwala'
    - 'ndiwodwala' -> 'ndi (w) odwala'

Chichewa





The word *odwala* ('patient') in 600 text-messages in Chichewa and the English translations

Chichewa

- Morphology: affixes and incorporation

  ndi-ta-ma-mu-**fun**-a-nso

  1PS-IMPLORE-PRESENT-2PS-**want**-VERB-also

  “I am also currently wanting you very much”

  a-ta-ma-ka-**fun**-a-nso

  class2.PL-IMPLORE-PRESENT-class12.SG-**want**-VERB-also

  “They are also currently wanting it very much”

- More than 30 forms for *fun* (‘want’), 80% novel

Haitian Krèyol

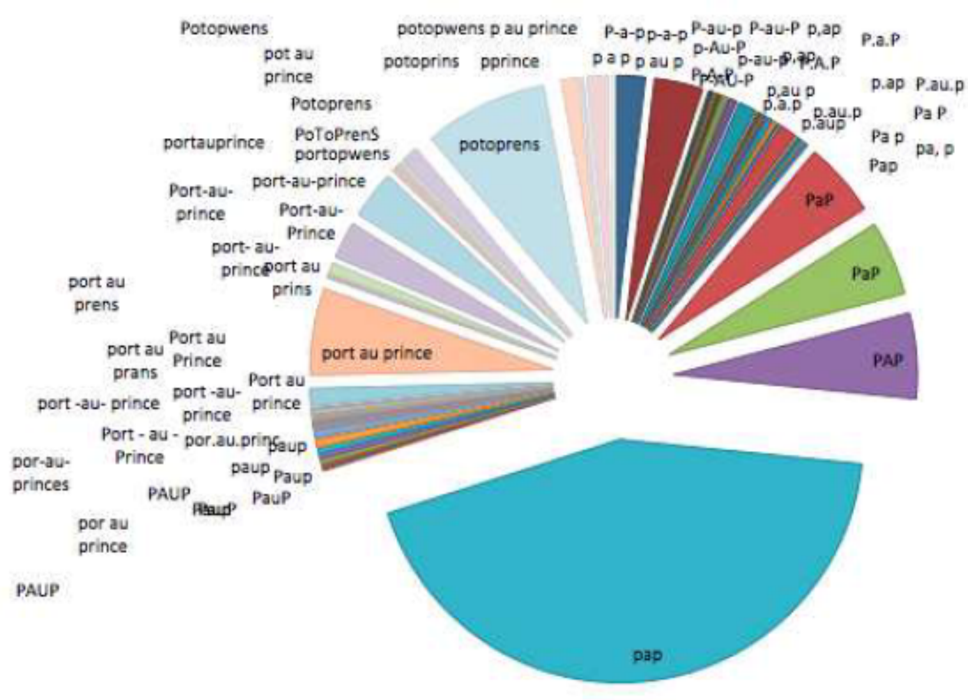
- More or less French spellings
- More or less phonetic spellings
- Frequent words (esp pronouns) are shortened and compounded
- Regional slang / abbreviations

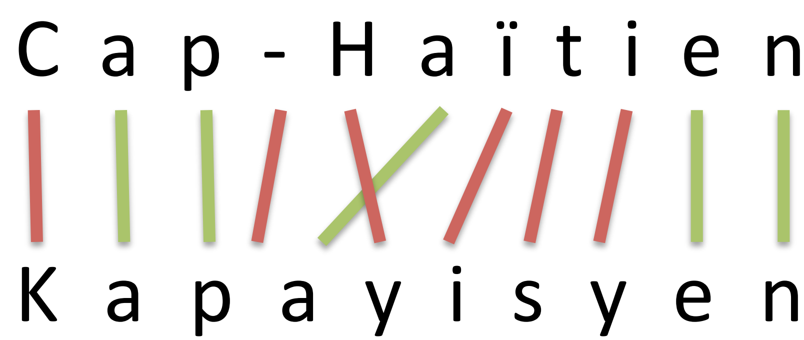
Haitian Krèyol

mèsi, mesi,

mèci, merci

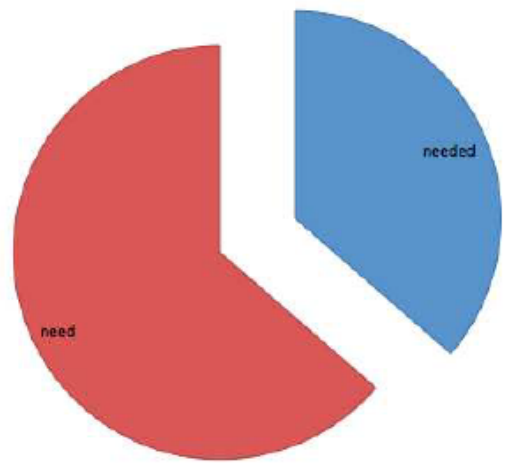
Abbrev.	**Full Form**	**Pattern**	**Meaning**
s'on	se yon	*sVn*	is a
avèn	avèknou	*VvVn*	with us
relem	rele mwen	*relem*	call me
wap	ouap	*uVp*	you are
map	mwen ap	*map*	I will be
zanmim	zanmi mwen	*zanmim*	my friend
lavel	lave li	*lavel*	to wash (it)





Urdu

- The least variant of the three languages here
  - Derivational morphology
    - *Zaroori / zaroorath*
  - Vowels and nonphonemic characters
    - *Zaruri / zaroorat*



zaroori ('need')

If it follows patterns, we can model it

Outline

- What do short message communications look like in most languages?
- How can we model the inherent variation?
- Can we create accurate classification systems despite the variation?
- Can we leverage loosely aligned translations for information extraction?

Subword models

- Segmentation
  - Separate into constituent morphemes:

    nditamamufunanso -> ndi-ta-ma-mu-fun-a-nso
- Normalization
  - Model phonological, orthographic, more or less phonetic spellings:

    odwela, edwala, odwara -> odwala

Language Specific

- Segmentation
  - Hand-coded morphological parser (Mchombo, 2004; Paas, 2005) [1]
- Normalization
  - Rule-based

    ph -> f, etc.

Linguistic paradigm	Form
Verb Prefixes and pre-Clitics:	
Negation	si
Subj Noun Classes	a, u, w, i, li, chi, zi, ka, ti, ku, pa, mu, ndi
Imperative	ta
Subjunctive modifiers	kana, kada
Tenses/Aspect	ku, ma, pa, dza, a, ba, ka
Negation	sa
Modals	nga, zi, ba, ta
Conditional	ka
Directives	dza, ka, dzi
2nd Modal	ngo
Obj Noun Classes	mu, wa, u, i, li, chi, zi, ka, ti
Verb Suffixes and post-Clitics	
Reciprocal	an
Causitive	its, ets
Applicative	il, el, i
Stative	ik, ek
Passive	idw, edw
Reversive	ul
Subjunctive	e
Final Vowel	a, i, o
Imperative	ni
Clitics	be, nso, tu, zi

Table 4.1: Morphological paradigms for Chichewa verbs

[1] robertmunro.com/research/chichewa.php

Language Independent

- Segmentation (Goldwater et al., 2009)
  - Context Sensitive Hierarchical Dirichlet Process, with morphemes, m_i drawn from distribution G generated from Dirichlet Process $DP(\alpha_0, P_0)$, with $H_m = DP$ for a specific morpheme:

$$\begin{aligned} m_i \mid m_{i-1} = m, H_m &\sim H_m && \forall m \\ H_m \mid \alpha_1, G &\sim DP(\alpha_1, G) && \forall m \\ G \mid \alpha_0, P &\sim DP(\alpha_0, P_0) \end{aligned}$$

- Extension to morphology:
  - Enforce existing spaces as morpheme boundaries
  - Identify free morphemes as min P_0, per word

ndi mafuna -> *ndi-ma-funa* *manthwala*

Language Independent

- Normalization
  - Motivated from minimal pairs in the corpus, *C*
  - Substitution, *H*, applied to a word, *w*, producing *w'* iff $w' \in C$

ndiwodwala -> ndiodwala

Form	Alternation
r([aeiouy])	l$1
([aeiou]\s*)[hwy]([aeiou])	$1$2
([a-z])\1+	$1
n([tdpbk])	$1
([tk]h)	$1
mn	n
sh	ch
c([aeiouy])	s$1
t	d
g	k
p	b
y	i
e	i
u	a
a	e
o	a
s	z

Table 4.3: Phonetically, phonologically & orthographically motivated alternation candidates.

Evaluation – downstream accuracy

- Most morphological parsers are evaluated on gold data and limited to prefixes or suffixes only:
  - *Linguistica* (Goldsmith, 2001), *Morphessor* (Creutz, 2006)
- Classification accuracy (macro-f, all labels):

	Chichewa Specific	*Language independent*
Segmentation:	**0.476**	0.425
Normalization:	0.396	**0.443**
Combined:	**0.484**	0.459

Other subword modeling results

- Stemming vs Segmentation
  - Stemming can *harm* Chichewa [1]
  - Segmentation most accurate when modeling discontinuous morphemes [1]
- Hand-crafted parser
  - Over-segments non-verbs (cf *Porter Stemmer* for English)
  - Under-segments compounds
- Acronym identification
  - Improves accuracy & can be broadly implemented [1]

[1] Munro and Manning, (2010)

Are subword models needed for classification?

Outline

- What do short message communications look like in most languages?
- How can we model the inherent variation?
- Can we create accurate classification systems despite the variation?
- Can we leverage loosely aligned translations for information extraction?

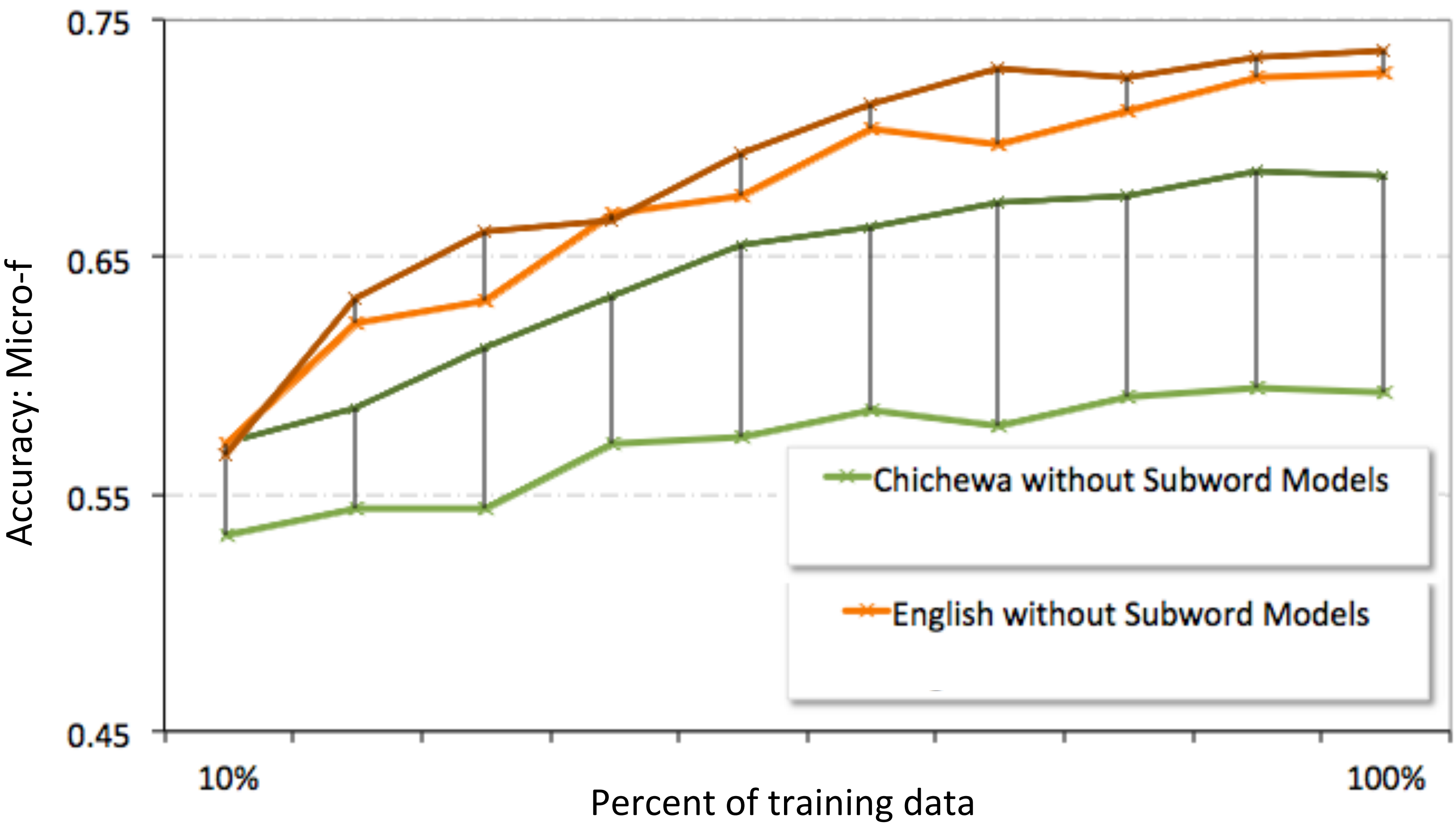
Classification

- Stanford Classifier
  - Maximum Entropy Classifier (Klein and Manning, 2003)
- Binary prediction of the labels associated with each message
  - Leave-one-out cross-validation
  - Micro-f
- Comparison of methods with and without subword models

Strategy

ndimmafuna manthwala ('I currently need medicine')	ndi kufuni mantwara ('my want of medicine')	1 in 5 classification errors with raw messages
ndimafuna mantwala	ndi kufuni mantwala	1) Normalize spellings
ndi-ma-fun-a man-twala	ndi-ku-fun-i man-twala	2) Segment
ndi-ma-**fun**-a **man-twala**	**ndi**-ku-**fun**-i **man-twala**	3) Identify predictors
ndi -fun man-twala ("I need medicine") Category = "Request for aid"	ndi -fun man-twala ("I need medicine") Category = "Request for aid"	1 in *20* classification error post-processing. *Improves* with scale.

Comparison with English

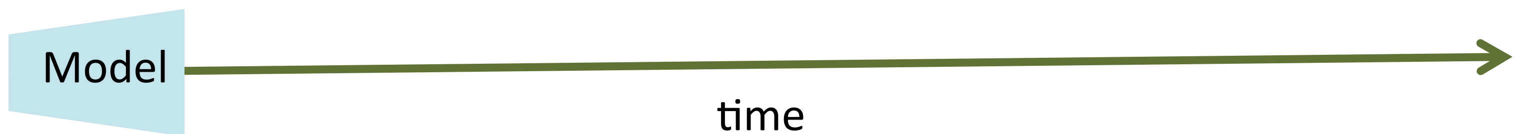


Streaming architecture

- Potential accuracy in a live, constantly updating system
  - Time sensitive and time-changing
- Kreyol ‘is actionable’ category
  - Any message that could be responded to
    (request for water, medical assistance, clustered requests for food, etc)

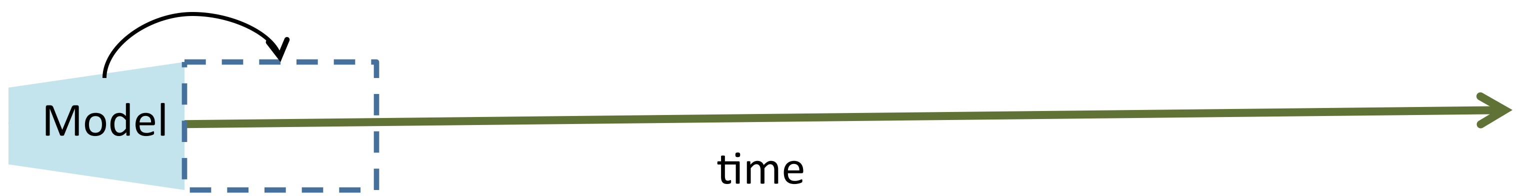
Streaming architecture

- Build from initial items



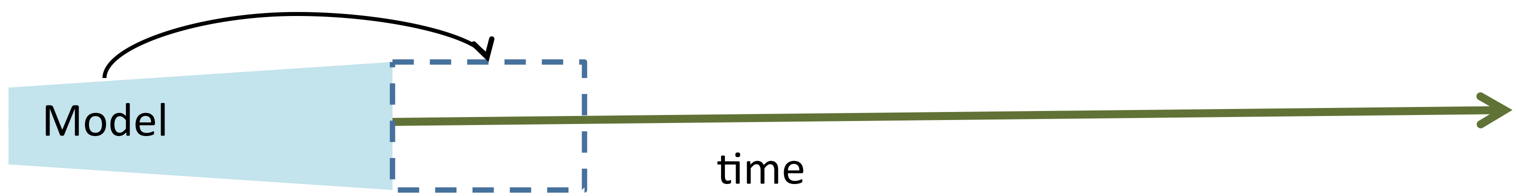
Streaming architecture

- Predict (and evaluate) on incoming items
  - (penalty for training)



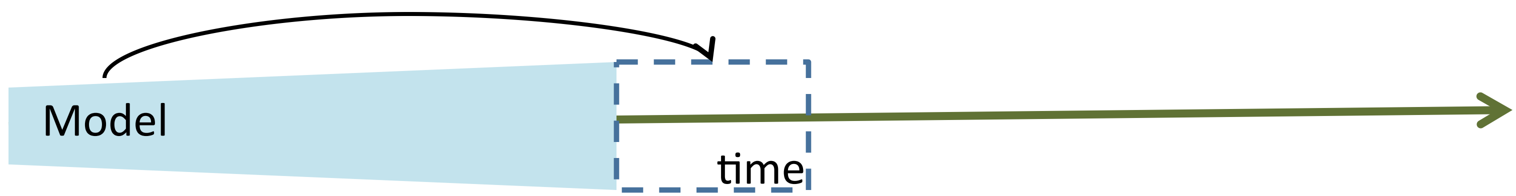
Streaming architecture

- Repeat / retrain



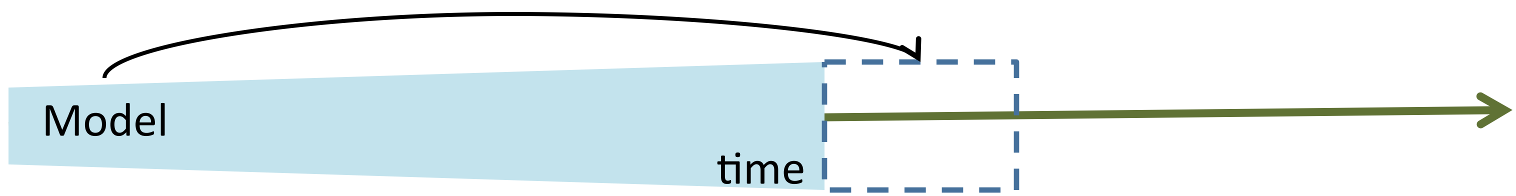
Streaming architecture

- Repeat / retrain



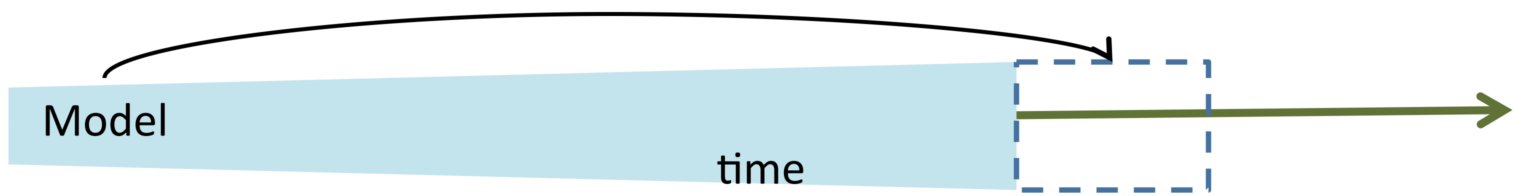
Streaming architecture

- Repeat / retrain



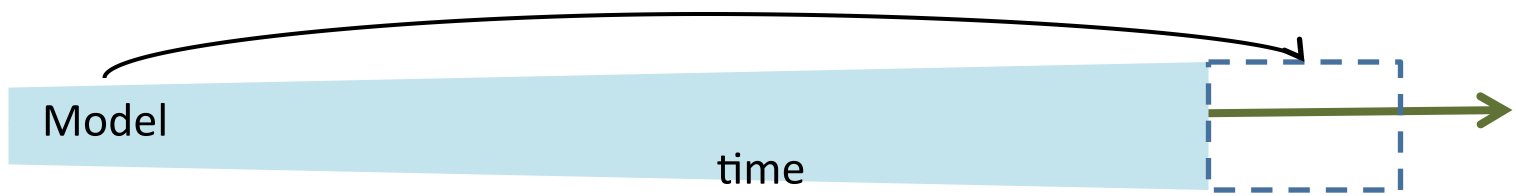
Streaming architecture

- Repeat / retrain



Streaming architecture

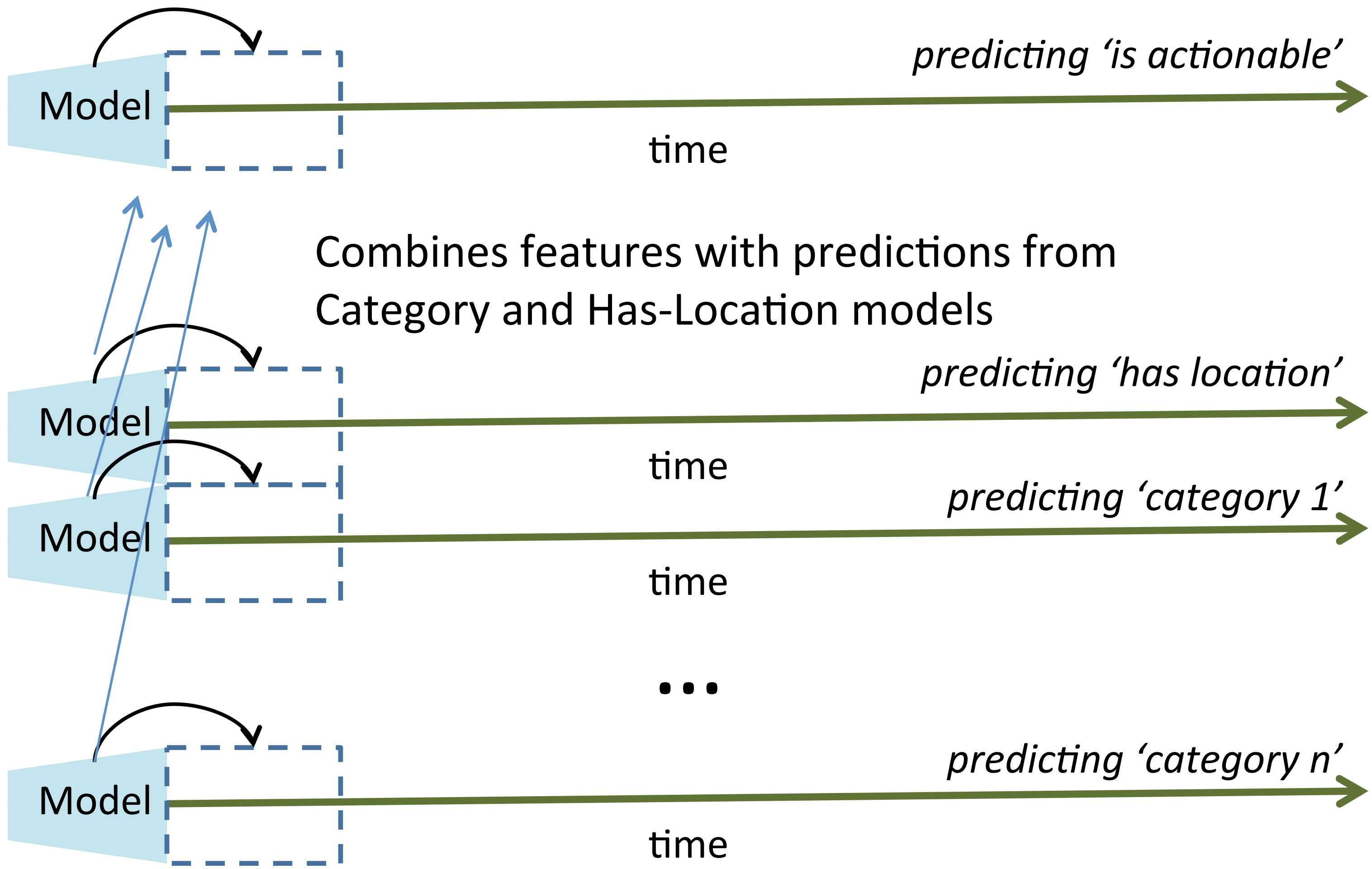
- Repeat / retrain



Features

- G : Words and ngrams
- W : Subword patterns
- P : Source of the message
- T : Time received
- C : Categories ($c_{0,\ldots,47}$)
- L : Location (longitude and latitude)
- $L_{\exists}$: Has-location (a location is written in the message)

Hierarchical prediction for 'is actionable'



Results – subword models

- Also a gain in streaming models

	Precision	Recall	F-value
Baseline	0.622	0.124	0.207
W Subword	0.548	0.233	**0.326**

Results – overall

- Gain of F > 0.6 for full hierarchical system, over baseline of words/phrases only

	Precision	Recall	F-value
Baseline	0.622	0.124	0.207
Final	0.872	0.840	**0.855**

Other classification results

- Urdu and English
  - Subword models improve Urdu & English tweets [1]
- Domain dependence
  - Modeling the source improves accuracy [1]
- Semi-supervised streaming models
  - Lower F-value but consistent prioritization [2]
- Hierarchical streaming predictions
  - Outperforms oracle for 'has location' [2]
- Extension with topic models
  - Improves non-contiguous morphemes [3]

[1] Munro and Manning, (2012); [2] Munro, (2011); [3] Munro and Manning, (2010)

Can we move beyond classification to information extraction?

Outline

- What do short message communications look like in most languages?
- How can we model the inherent variation?
- Can we create accurate classification systems despite the variation?
- Can we leverage loosely aligned translations for information extraction?

Named Entity Recognition

- Identifying mentions of People, Locations, and Organizations
  - Information extraction / parsing / Q+A
- Typically a high-resource task
  - Tagged corpus (Finkel and Manning, 2010)
  - Extensive hand-crafted rules (Chiticarui, 2010)
- How far can we get with loosely aligned text?
  - One of the only resources for most languages

Example

Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid 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 Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

The intuition

Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid 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 Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

Do named entities have the least edit distance?

The intuition

Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid 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 Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

The intuition

Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid 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 Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

The intuition

Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid 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 Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

The intuition

Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid 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 Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

The complications

Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid 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 Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.

Capitalization of entities was not always consistent

Slang/abbreviations/alternate spellings for 'Okap' are frequent: 'Cap-Haitien', 'Cap Haitien', 'Kap', 'Kapayisyen'

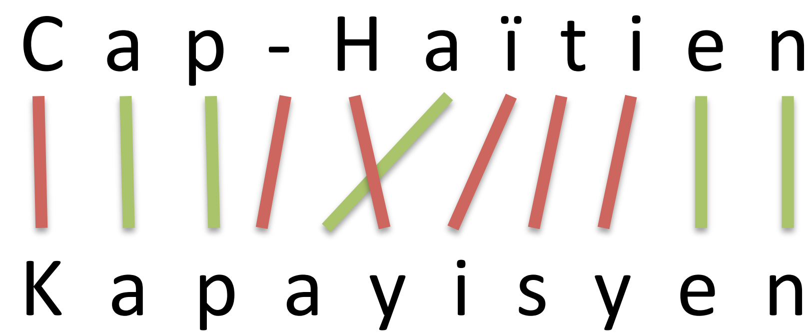
3 Steps for Named Entity Recognition

1. Generate seeds by calculating the edit likelihood deviation.
2. Learn context, word-shape and alignment models.
3. Learn weighted models incorporating supervised predictions.

Step 1: Edit distance (Levenshtein)

- Number of substitutions, deletions or additions to convert one string to another
  - *Minimum Edit Distance:* min between parallel text
  - *String Similarity Estimate:* normalized by length
  - *Edit Likelihood Deviation:* similarity, relative to average similarity in parallel text (z-score)
  - *Weighted Deviation Estimate:* combination of Edit Likelihood Deviation and String Similarity Estimate

Example



- *Edit distance*: 6
- *String Similarity*: ~0.45

“Voye manje medikaman pou moun kie nan lopital Kapayisyen”

“Send food and medicine for people in the Cap Haitian hospitals”

- Average & standard dev similarity: μ=0.12, σ=0.05
- *Edit Likelihood Deviation:* 6.5 (good candidate)

“Voye manje medikaman pou moun kie nan lopital Kapayisyen”

“They said to send manje medikaman for lopital Cap Haitian”

- Average & standard dev similarity: μ=0.21, σ=0.11
- *Edit Likelihood Deviation:* 2.2 (doubtful candidate)

Equations for edit-distance based metrics

- Given a string in a message and translation $M_{S,}\ M'_{S'}$

Levenshtein distance LEV()

String Similarity Estimate SSE()

$$SSE(M_S, M'_{S'}) = 1 - \frac{(2(LEV(M_S, M'_{S'})) + 1}{LEN(M_S) + LEN(M'_{S'}) + 1}$$

Average AV()

Standard Deviation SD()

Edit Likelihood Deviation ELD()

$$ELD(M_S, M'_{S'}) = \frac{(SSE(M_S, M'_{S'})) - AV(SSE(M_{0-n}, M'_{0-m}))}{SD(SSE(M_{0-n}, M'_{0-m}))}$$

Normalizing Function N()

Weighted Deviation Estimate WDE()

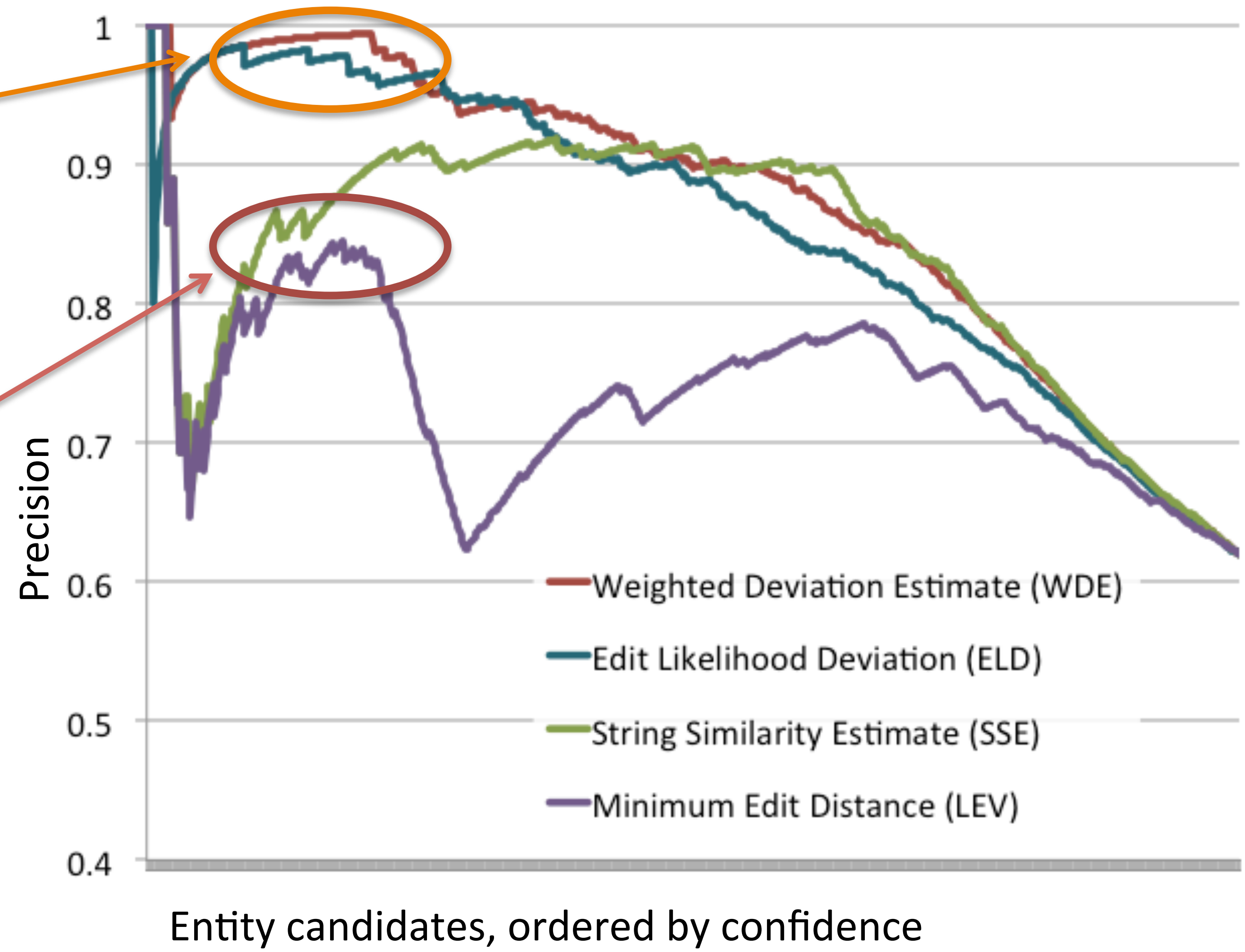
$$WDE(M_S, M'_{S'}) = (SSE(M_S, M'_{S'})^{\alpha}.N(ELD(M_S, M'_{S'})^{1-\alpha}))^2$$

Comparison of edit-distance based metrics

Novel to this research: local deviation in edit-distance.

Past research used global edit-distance metrics (Song and Strassel, 2008)

This line of research not pursued after *REFLEX* workshop.



Step 2: Seeding a model

- Take the top 5% matches by WDE()
  - Assign an 'entity' label
- Take the bottom 5% matches by WDE()
  - Assign a 'not-entity' label
- Learn a model
- Note: the bottom 5% were still the best match for the given message/translation
  - Targeting the boundary conditions

Features

… ki nan vil Milot, 14 km nan sid …

… located in this village Milot 14 km south of …

- Context: BEF_vil, AFT_14 / BEF_village, AFT_14
- Word Shape: SHP_Ccp / SHP_Cc
- Subword: SUB_<b>Mi, SUB_<b>Mil, SUB_il, …
- Alignment: ALN_8_words, ALN_4_perc
- Combinations: SHP_Cc_ALN_4_perc, …

Strong results

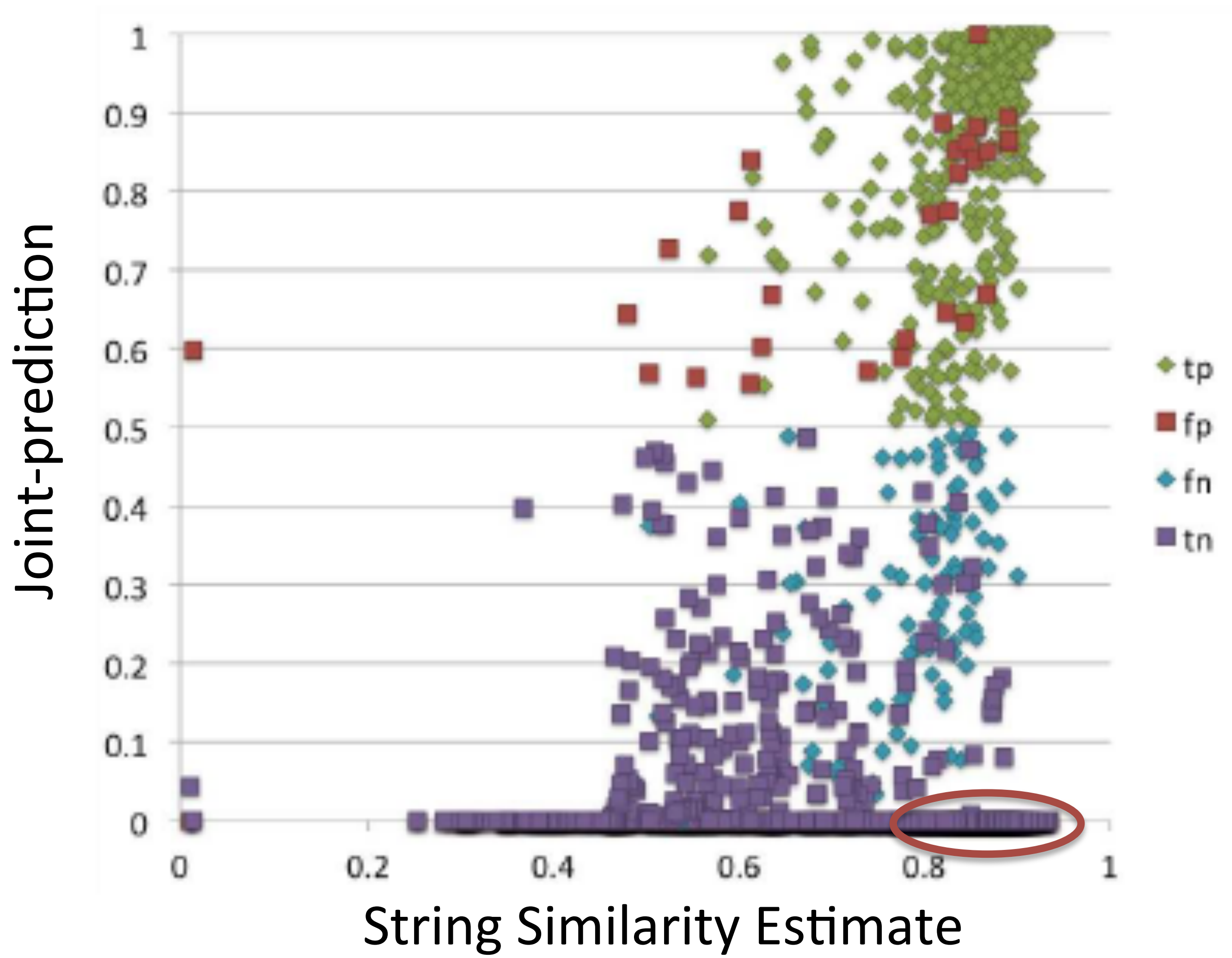
- Joint-learning across both languages

	Precision	Recall	F-value
Kreyol	0.904	0.794	0.846
English	0.915	0.813	**0.861**

- Language-specific:

	Precision	Recall	F-value
Kreyol	0.907	0.687	0.781
English	0.932	0.766	0.840

Effective extension over edit-distance



Domain adaption

Completely unsupervised, using ~3,000 sentences loosely aligned with Kreyol

- Joint-learning across both languages

	Precision	Recall	F-value
Kreyol	0.904	0.794	0.846
English	0.915	0.813	0.861

- Supervised (MUC/CoNLL-trained Stanford NER):

	Precision	Recall	F-value
English	0.915	0.206	0.336

Fully supervised, trained over 10,000s of manually tagged sentences in English

Step 3: Combined supervised model

… ki nan vil Milot, 14 km nan sid …

… located in this village Milot 14 km south of …

Step 3a: Tag English sequences from a model trained on English corpora (Sang, 2002; Sang and De Meulder, 2003; Finkel and Manning, 2010)

Step 3b: Propagate across the candidate alignments, in combination with features (context, word-shape, etc)

Combined supervised model

- Joint-learning across both languages

	Precision	Recall	F-value
Kreyol	0.904	0.794	0.846
English	0.915	0.813	0.861

- Combined Supervised and Unsupervised

	Precision	Recall	F-value
Kreyol	0.838	0.902	0.869
English	0.846	0.916	0.880

Other information extraction results

- Other edit-distance functions (eg: Jaro-Winkler)
  - Make little difference in the seed step - the deviation measure is the key feature [1]
- Named entity discrimination
  - Distinguishing People, Locations and Organizations is reasonably accurate with little data [1]
- Clustering contexts
  - No clear gain – probably due to sparse data

[1] Munro and Manning, (*under review*)

Outline

- What do short message communications look like in most languages?
- How can we model the inherent variation?
- Can we create accurate classification systems despite the variation?
- Can we leverage loosely aligned translations for information extraction?

Conclusions

- It is necessary to model the subword variation found in many of the world's short-message communications
- Subword models can significantly improve classification tasks in these languages
- The same subword variation, cross-linguistically, can be leveraged for accurate named entity recognition

Conclusions

- More research is needed

2012 (estimate): 9 Trillion

2007 (start of PhD): 5 Trillion

2000: 1 Trillion

Thank you

References

Munro, R. and Manning, C. D. (2010). Subword variation in text message classification. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, Los Angeles, CA.

Munro, R. (2011). Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol. *Proceedings of the Fifteenth Conference on Natural Language Learning (CoNLL 2011)*, Portland, OR.

Munro, R. and Manning, C. D. (2012). Short message communications: users, topics, and in-language processing. *Proceedings of the Second Annual Symposium on Computing for Development (ACM DEV 2012)*, Atlanta, GA.

Munro, R. and Manning, C. D. (*2012*). Accurate Unsupervised Joint Named-Entity Extraction from Unaligned Parallel Text. *Proceedings of the Named Entities Workshop (NEWS 2012)*, Jeju, Korea.