Introductory Course at ESSLLI

Bolzano, Italia

August 2016

Crowdsourcing Linguistic Datasets

LECTURE 4

Chris Biemann

biem@cs.tu-darmstadt.de

Lesson 4: A Few Sample Projects

Projects were selected to span a wide range of NLP tasks and whether they discuss crowdsourcing-related issues

Coming up next:

- Part-of-Speech tagging
- Named Entity Recognition and Classification
- Prepositional Phrase Attachment
- Word Alignment
- Relation Extraction
- Question Rating
- Image Annotation

Since crowdsourcing has become a commodity, there are less and less papers that specifically discuss crowdsourcing practices for NLP.

Many examples from the NAACL 2010 workshop on Creating Speech and Language Data with Amazon's Mechanical Turk: 24 Participants were granted $100 each to promote crowdsourcing in NLP

Use of Manually Acquired Data in NLP

- Resource Creation
  - putting together a dictionary for human or machine use
  - includes: Wikipedia, Wiktionary, WordNet
- Training Data Acquisition
  - create training / development / test data for machine learning
  - includes: treebanking, text annotation, translation, document class labeling, marking as spam
- Evaluation
  - have system output manually checked
  - post-hoc evaluations for all sorts of NLP systems

All of the above can be crowdsourced, but pose different challenges – mostly related to the missing expertise of the average crowdworker, as well as quality control in light of the vagueness/variety of language.

Crowdsourced Re-annotation of POS tagging data

- Task: assign parts of speech to Twitter data
  Q/**NOUN** :/. hay/**PRT** justin/**NOUN** SCREEEEEEEEM/**PRT** !!!!!!!!/. i/**PRON** luv/**VERB** u/**PRON** OMG/**PRT** !!!!!!!!!/. i/**VERB** did/**VERB** a/**DET** quiz/**NOUN** ubout/**ADP** if/**ADP** me/**PRON** and/**CONJ** u/**PRON** wer/**VERB** thu/**DET** only/**ADJ** ones/**PRON** o/**ADP** http://www.society.me/q/29910/view/**X**
- Motivation: Language change on Twitter is rapid, thus models fall out of use quickly

Crowdflower Interface

Hovy, D., Plank, B., Søgaard, A. (2014): Experiments with crowdsourced re-annotation of a POS tagging data set. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 377–382, Baltimore, MD, USA

Crowdsourced Re-annotation of POS tagging data II

- Crowd Setup on Crowdflower:
  - only trusted crowdworkers: need to pass 4 test items
  - reward: $0.05 for 10 tokens / 5 annotations per token, thus 2.5 cents / token
  - full dataset: 14,619 tokens, took 10 days to complete
  - high satisfaction of crowdworkers with the task
- Aggregation: comparing Majority Voting (MV) with MACE
  - MV: treat all annotators equally and choose the label that most annotators supply
  - MACE: treat annotator competence and true label as hidden variables and estimate both with Expectation Maximization (Hovy et al., 2013)
- Evaluation:
  - compare to gold standard labels from expert annotators
  - compare ML model quality
  - compare impact on a downstream tasks, here: chunking and NER

Hovy, D. Berg-Kirkpatrick, T., Vaswani, A., Hovy, E. (2013). Learning whom to trust with MACE. Proceedings of NAACL-2013, Atlanta, GA, USA.

Crowdsourced Re-annotation of POS tagging data III

- over 10% of tokens never received gold label, mostly related to punctuation and pronouns
- MACE scheme helps a little, filtering with Wiktionary helps more
- impact on downstream: yes for chunking, no for NER

majority	79.54
MACE-EM	79.89
majority+Wiktionary	80.58
MACE-EM+Wiktionary	80.75
oracle	89.63

Table 1: Accuracy (%) of different annotations wrt gold data

x	Z	y
@USER	NOUN,NOUN,X,NOUN,–,NOUN	NOUN
:	.,.,–,.,.,.	X
I	PRON,NOUN,PRON,NOUN,PRON,–	PRON
owe	VERB,VERB,–,VERB,VERB,VERB	VERB
U	PRON,X,–,NOUN,NOUN,PRON	PRON

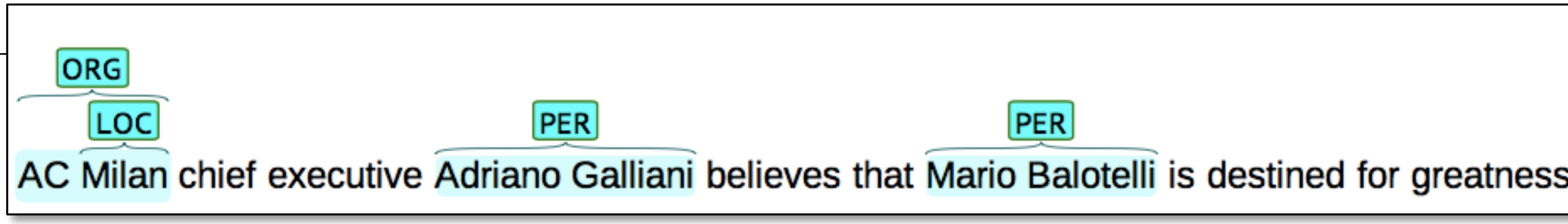
$$\theta = 0.9, 0.4, 0.2, 0.8, 0.8, 0.9$$

Figure 1: Five annotations per token, supplied by 6 different annotators (– = missing annotation), gold label $\mathbf{y}$. θ = competence values for each annotator.

POS model from	CHUNKING	NER
MV	74.80	75.74
MACE	75.04	75.83
MV+Wik	75.86	76.08
MACE+Wik	75.86	76.15
Upper bounds		
oracle	76.22	75.85
gold	79.97	75.81

Table 3: Downstream accuracy for chunking (l) and NER (r) of models using POS.

Named Entity Recognition with Crowdsourcing



- Task: mark name spans in text and assign a class label
- Challenge for crowdsourcing:
  - standard interface does not support the marking of spans
  - payment scheme encourages low recall if we pay ‘per paragraph’
- Solution:
  - custom interface
  - bonus system using command line tools

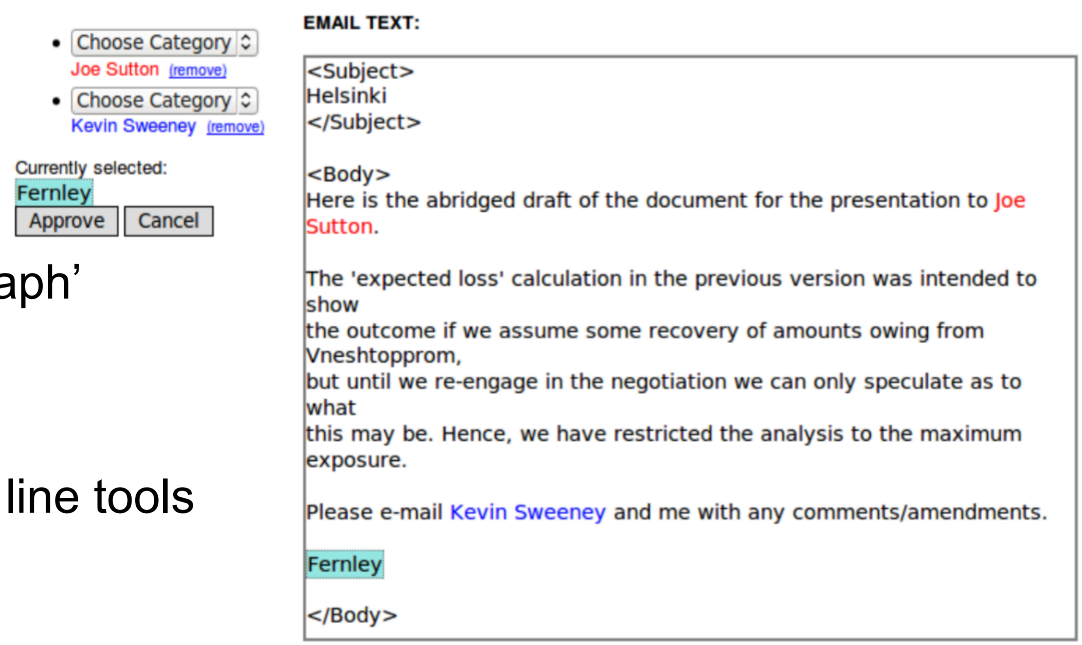


Fig. 1: Sample of the interface presented to workers.

Lawson, N., Eustice, K., Perkowitz, M. (2010): Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, pages 71–79, Los Angeles, CA, USA

Named Entity Recognition with Crowdsourcing II

- Web-based GUI that supports highlighting/marking of tokens, written in JavaScript
- Annotation of 20,609 email messages of 400 characters on average
- looking for three types PERson, ORGanization and LOCation separately: in each task, only one type is sought for
  - for PER, workers also annotated unnamed mentions like “my mom”, thus a separate class of these was included, just to discard its contents for NER
- Pricing scheme on Amazon MTurk
  - $0.01 for each HIT – regardless of the number of entities found
  - $0.01 / $0.02 bonus for each entity found
  - Bonus only paid if the majority of annotators found the respective entity
- Setup on MTurk
  - batches of 100 – 1000 emails: larger batches completed faster
  - 798 workers in total, only 10 scammers that never marked any entity

Named Entity Recognition with Crowdsourcing III

- Different types have different recall levels: need more workers to catch all LOCs and ORGs, fewer to catch PERs
- Bonus system seems to work: most productive workers tend to have a high recall
- using annotations that at least 2 workers marked produced best tagging results (more: recall too little; less: precision issues)

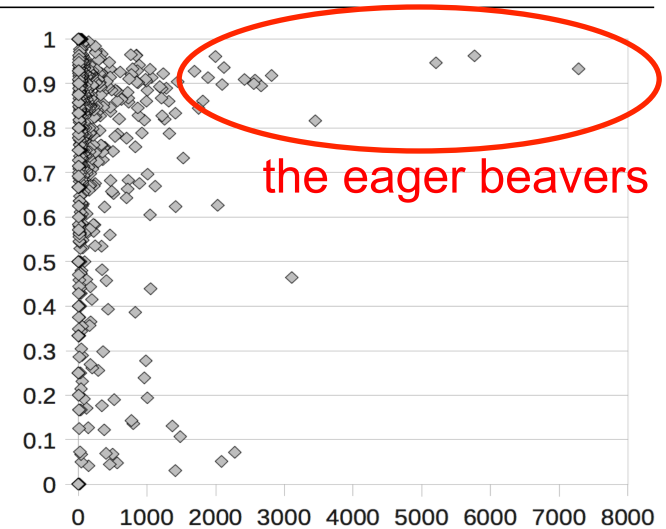


Fig. 4: # HITs Completed vs. Recall

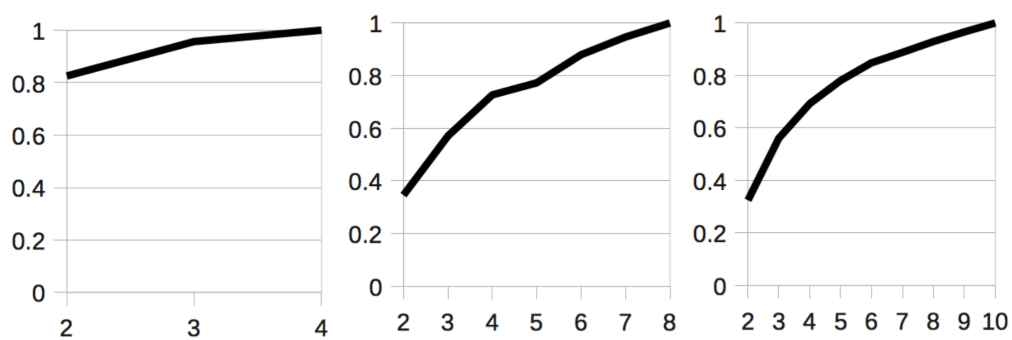


Fig. 3: Marginal recall curves for PERSON, LOCATION, and ORGANIZATION entity types, from a trial run of 900-1,000 emails. Recall is plotted on the y-axis, the number of annotators on the x-axis.

Named Entity Recognition with Crowdsourcing IV

- Alternative interface using standard forms (generated per HIT)
- more complex, does not handle overlapping annotations
- was tested only on small batches, hence unclear how scammers should be handled when scaling up

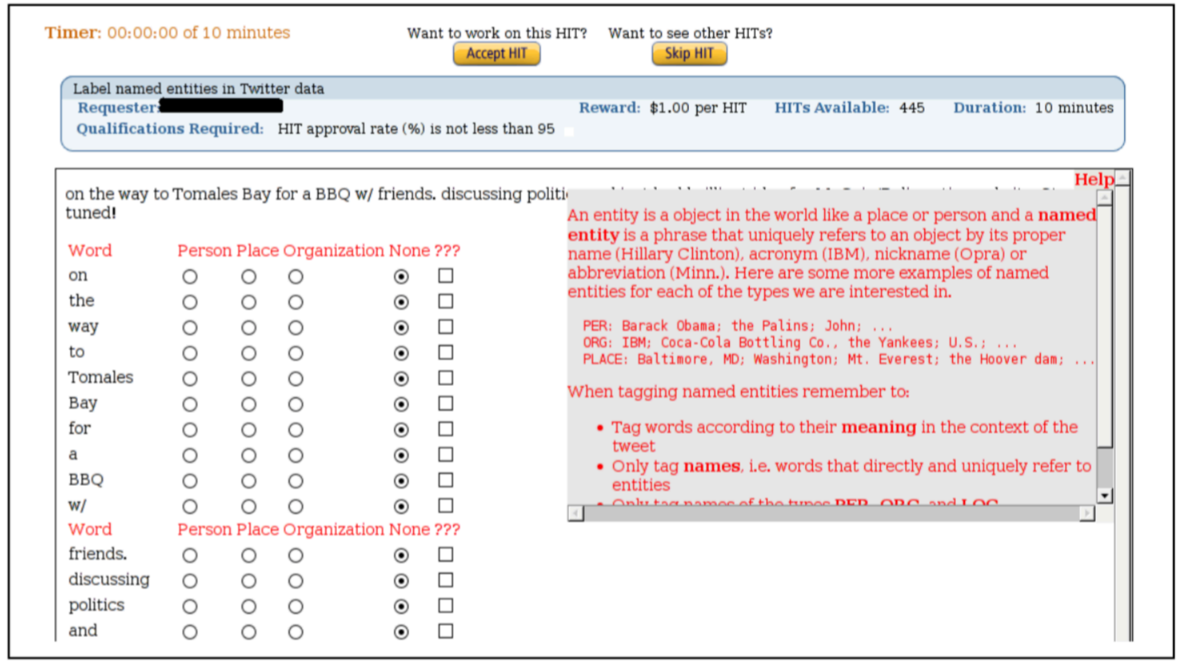


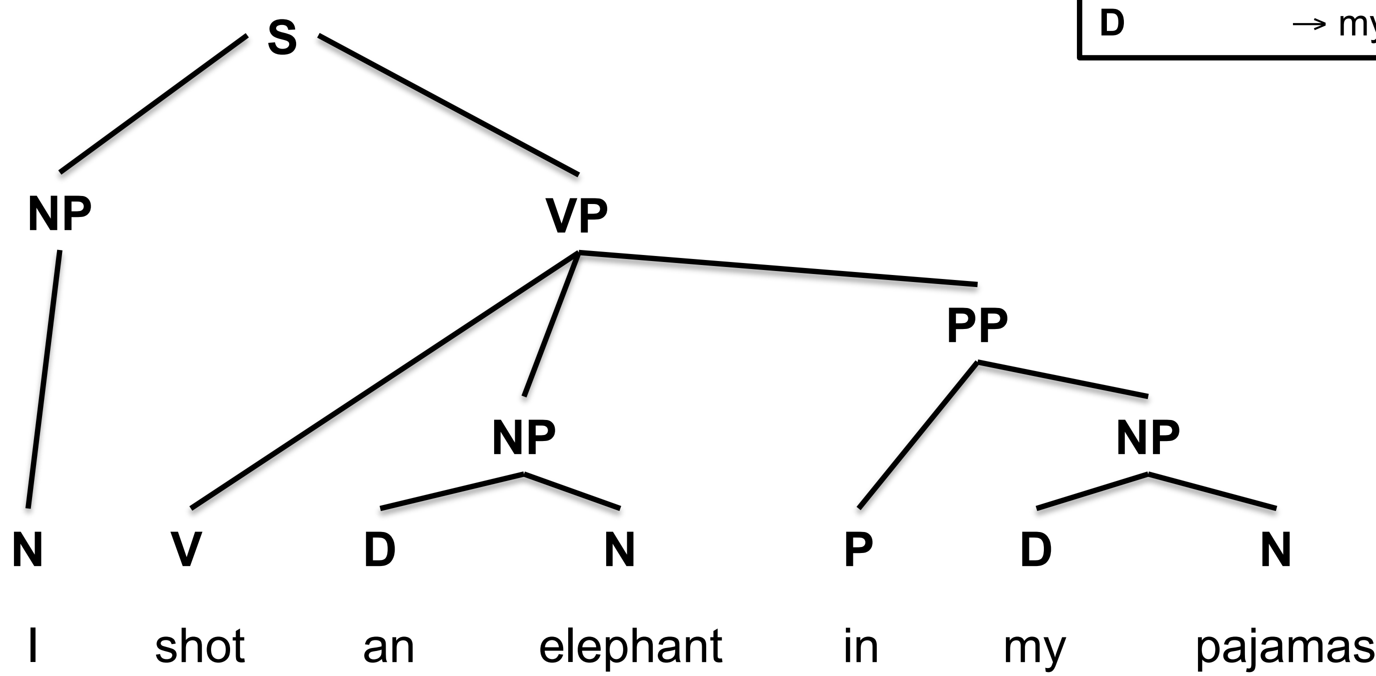
Figure 3: In the MTurk interface a tweet is shown in its entirety at the top, then a set of radio buttons and a checkbox is shown for each word of the tweet. These allow the user to pick the annotation for each word, and indicate uncertainty in labeling.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M. (2010): Annotating Named Entities in Twitter Data with Crowdsourcing. Procee–dings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 80–88, Los Angeles, CA, USA

PP Attachment: Major Issue for Phrase Structure Grammars

- Syntax trees can (almost) be modeled with context-free languages

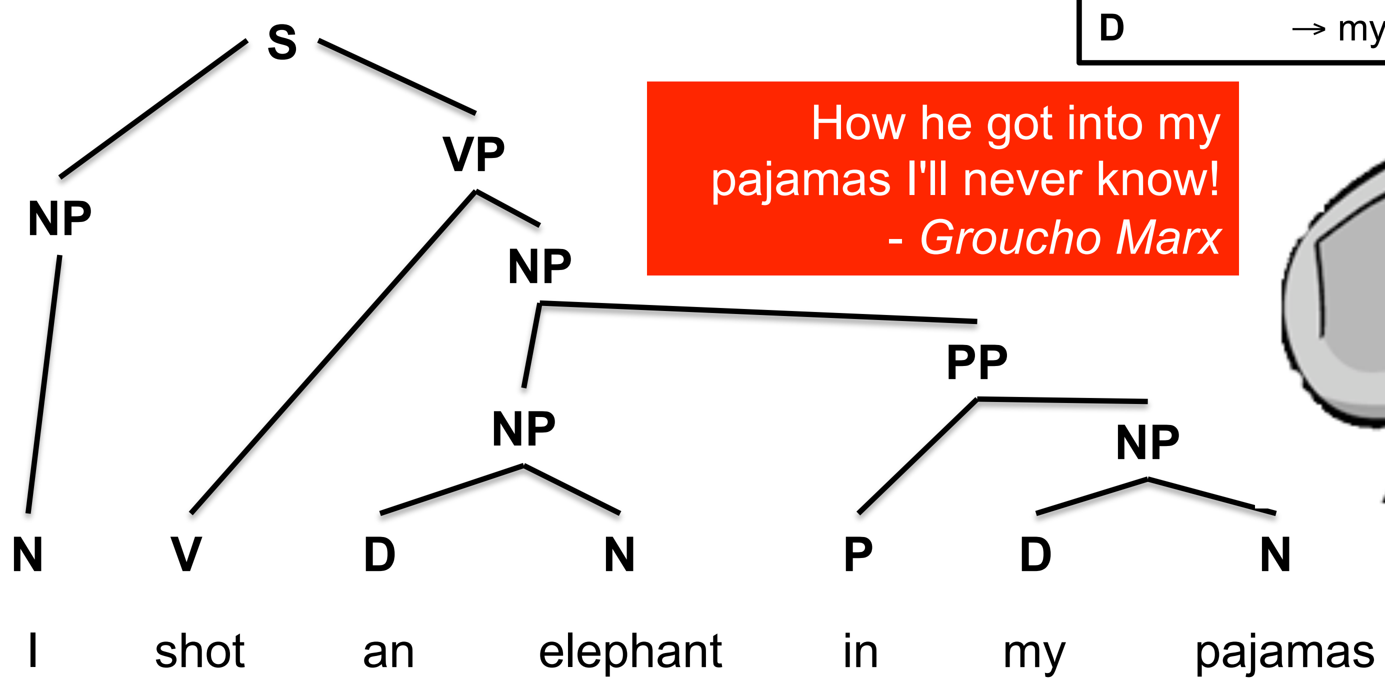
S	**→ NP VP**
NP	**→ N | D N | NP PP**
VP	**→ V NP | V NP PP**
PP	**→ P NP**
N	→ I | elephant | pajamas
V	→ shot
P	→ in
D	→ my | an | a | the



PP Attachment: Major Issue for Phrase Structure Grammars

- Syntax trees can (almost) be modeled with context-free languages
- one surface sentence can have several derivations

S	**→ NP VP**
NP	**→ N | D N | NP PP**
VP	**→ V NP | V NP PP**
PP	**→ P NP**
N	→ I | elephant | pajamas
V	→ shot
P	→ in
D	→ my | an | a | the



How he got into my pajamas I'll never know!
\- *Groucho Marx*

Crowdsourcing for PP Attachment I

- Motivation: PP attachment bias is different for different genres
- Need semantic knowledge to disambiguate PP attachment ambiguities
- Setup:
  - generate possible attachments from POS tag sequences and chunks
  - generate crowdsourcing questions to decide the correct attachment

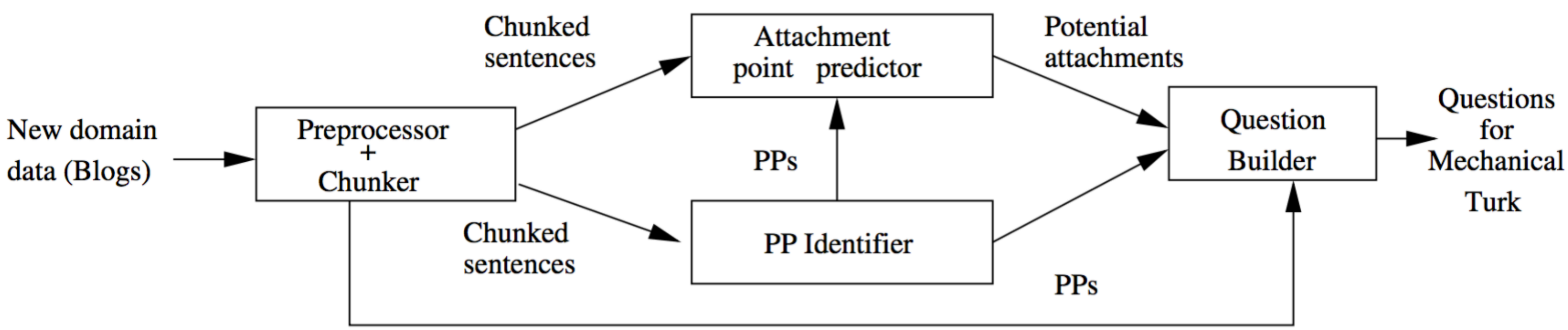


Figure 1: Overview of question generation system

Jha, M., Andreas, J., Thadani, K., Rosenthal, S., McKeown, K. (2010): Corpus Creation for New Genres: A Crowdsourced Approach to PP Attachment. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 13–20, Los Angeles, CA, USA

Crowdsourcing for PP Attachment II

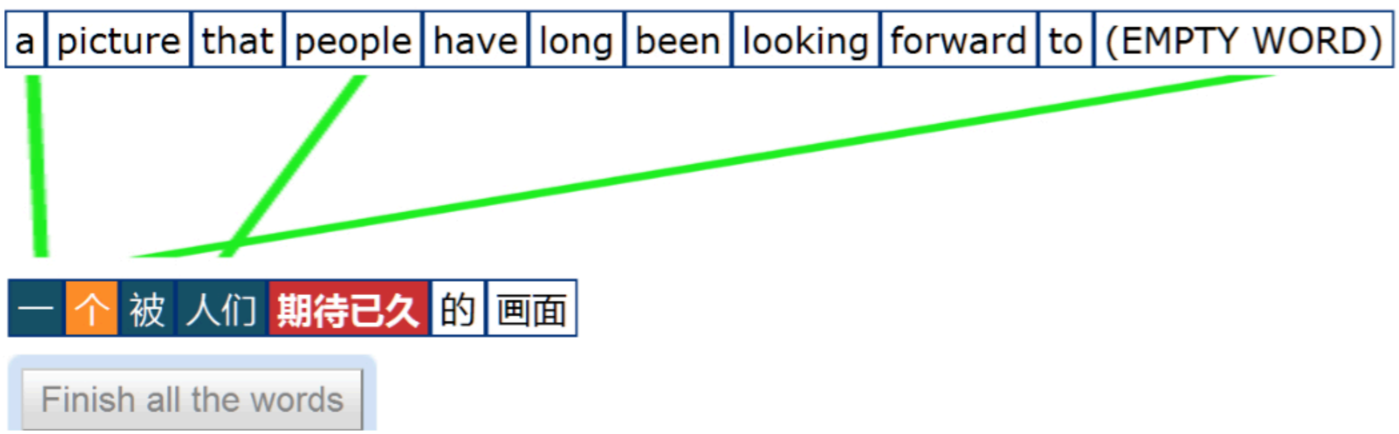
- Crowdsourcing Setup on MTurk:
  - show sentence with PP highlighted, allow to pick best option to attach it
  - exits: workers can type additional options, indicate problems with HIT
  - 1000 HITs, 5 workers per HIT, $0.04 per question
- Results
  - typical accuracy/multiplicity tradeoff
  - about 5% loss due to chunker errors – these were often identified with the "exit" option

Workers in agreement	Number of questions	Accuracy	Coverage
5 (unanimity)	389	97.43%	41.33%
$\geq$ 4 (majority)	689	94.63%	73.22%
$\geq$ 3 (majority)	887	88.61%	94.26%
$\geq$ 2 (plurality)	906	87.75%	96.28%
Total	**941**	**84.48%**	**100%**

Table 2: Accuracy and coverage over agreement thresholds

Crowdsourcing Word Alignment

- Motivation
  - Machine Translation systems learn from parallel data, usually from parallel sentences
  - word alignment is usually done automatically, but results in noise
- Solution: use crowdsourcing for word alignment
- Specialized interface on top of Google Web Kit (JavaScript)



Gao, Q. and Vogel, S. (2010): Consensus versus Expertise : A Case Study of Word Alignment with Mechanical Turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 30–34, Los Angeles, CA, USA

Crowdsourcing Word Alignment II

- Collecting and accepting alignments with majority vote leads to partial alignments in presence of worker noise
- Information from partial alignments: a) we get pairs of aligned words and b) we know which words they are NOT aligned to

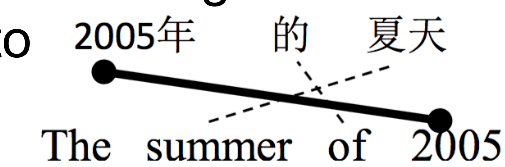


Figure 2: Partial and full alignments

- Using this information to constrain an automatic aligner reduces overall alignment error
- Other observation: lack of Chinese-speaking crowdworkers: task went slow, even after raising the price considerably.

	Partial	Full	Full-Int
Number of sentences	135	239	135
Number of words	2,008	3,241	2,008
Consensus words	13,03	2,299	1,426
Consensus rate(%)	64.89	70.93	71.02
Total Links	7,508	9,767	6,114
Consensus Links	5,625	7,755	4,854
Consensus Rate(%)	74.92	79.40	79.39
Total Unique Links	3,186	3,989	2,506
Consensus Links	1,875	2,585	1,618
Consensus Rate(%)	58.85	64.80	64.54
In majority group	2,447	3,193	1,426
Majority rate(%)	76.80	80.04	71.06

Table 1: Internal consistency of manual alignments, here Full-Int means statistics of full alignment tasks on the sentences that also aligned using partial alignment task

Crowds 4 Relation Extraction

1. The sentence expresses the relation.
   Sentence: For the past eleven years, James has lived in Tucson.
   Relation: "Tucson" is the residence of "James"
2. The sentence does not express the relation.
   Sentence: Samuel first met Divya in 1990, while she was still a student.
   Relation: "Divya" is a spouse of "Samuel"
3. The relation does not make sense.
   Sentence: Soojin was born in January.
   Relation: "January" is the birth place of "Soojin"

Figure 1: The three annotation options with examples.

- Motivation
  - relation annotation (e.g. born in, plays for ..) in text is expensive
  - distant supervision: use a knowledge base to find patterns in which known relations occur helps but is error-prone
  - can use crowdsourcing to manually correct wrong extractions
- Setup
  - show 10 sentences with relations (from 17 relations between persons) and have crowdworkers assign one of three options above
  - 7 are automatically generated, 3 control items
  - $0.05 per HIT, 5 workers/HIT

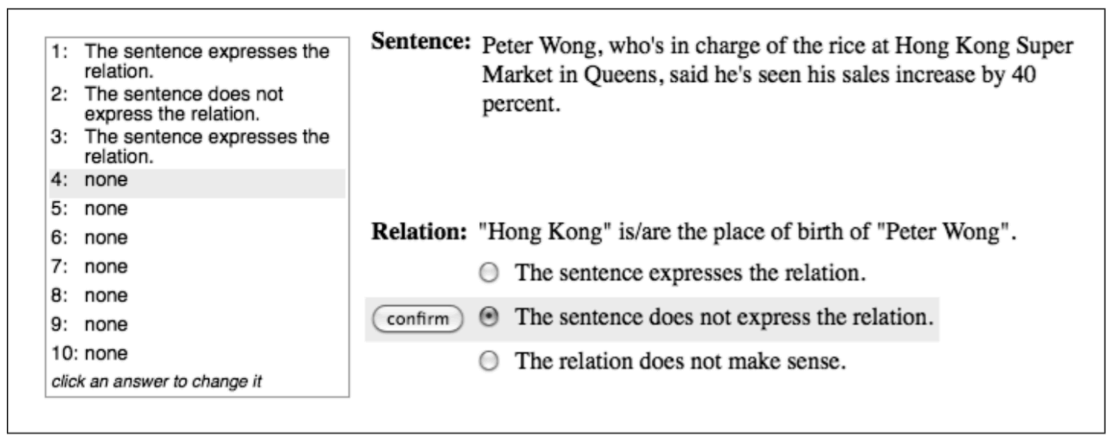


Figure 2: An example HIT with instructions excluded.

Gormley, M.R., Gerber, A., Harper, M., Dredze, M. (2010): Non-Expert Correction of Automatically Generated Relation Annotations. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 204–207, Los Angeles, CA, USA

Crowds 4 Relation Extraction II

- Inter-Annotator-Agreement: measures how much people provide the same labels for the same task. Commonly used: Cohen's Kappa
- Agreement often perceived as an upper bound for learning algorithms
- Here: expert annotators (E1/E2) show higher agreement than expert vs. majority vote (M); control questions seem "easier"
- Filtering bad workers: by control items and by time (too short is bad)

	# Ex.	R	Exact-κ	Pairwise
E1/E2	247	2	0.64	0.81
E1/M	247	2	0.29	0.60
E2/M	247	2	0.39	0.70
C/M	1059	2	0.90	0.93
T(sample)	247	5	0.31	0.69
T(control)	1059	5	0.52	0.68
T(all)	3530	5	0.45	0.68

Table 2: Inter-annotator agreement

	E1/M	*E2/M*
Unfiltered	0.28	0.38
Time Filtered	0.32	0.43
Control Filtered	0.34	0.47
Control and Time	0.37	0.48

Table 5: Exact-κ scores for three levels of quality control and a baseline, between each expert and the majority vote

Conger, A.J. (1980): Integration and generalization of kappas for multiple raters. Psychological Bulletin, 88(2):322–328.
Landis, J. R. and Koch, G. G. (1977): The measurement of observer agreement for categorical data. Biometrics, 33(1):159-74.

Question Rating with Crowdsourcing

- Goal: automatically generate reading comprehension questions
- Why? Because authors work for the Educational Testing Service – hands up: who participated in: GRE? TOEFL? PISA?
- Approach: overgenerate-and-rank paradigm: generate as many questions as possible, then pick the 'best' by statistical ranking
- Ranker (any ranker!) needs to be trained on manually judgments

	Rating	Details
1	Bad	The question has major problems.
2	Unacceptable	The question definitely has a minor problem.
3	Borderline	The question might have a problem, but I'm not sure.
4	Acceptable	The question does not have problems.
5	Good	The question is as good as one that a human teacher might write for a reading quiz.

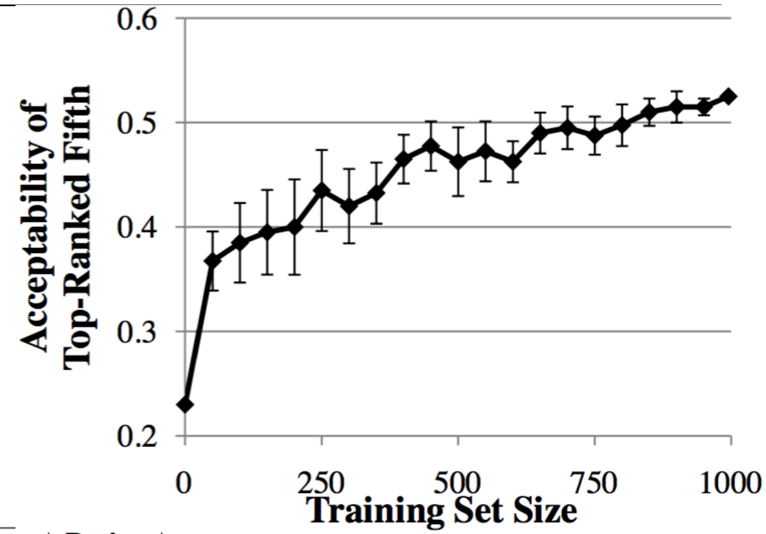
Table 1: The five-point question rating scale.

- Setup on Mturk:
  - $0.05 per rating, 5 workers/HIT
  - hourly wage: $5-$10 / hour
  - using default qualifications and manual filtering of bad workers

Heilman, M., Smith, N.A. (2010): Rating Computer-Generated Questions with Mechanical Turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 35-40, Los Angeles, CA, USA

Question Rating with Crowdsourcing II

- Results: averaging over 3-7 crowdworkers achieves the performance of a computational linguist, as measured by ranking correlation
- When using this data for training (linear regression on a set of 326 numerical features), data shows a very positive trend



Source Text Excerpt	Question	Rating
MD 36 serves as the main road through the Georges Creek Valley, a region which is historically known for coal mining, and has been designated by MDSHA as part of the Coal Heritage Scenic Byway.	*Which part has MD 36 been designated by MDSHA as?*	1.4
He worked further on the story with the Soviet author Isaac Babel, but no material was ever published or released from their collaboration, and the production of Bezhin Meadow came to an end.	*What did the production of Bezhin Meadow come to?*	2.0
The design was lethal, successful and much imitated, and remains one of the definitive weapons of World War II.	*Does the design remain one of the definitive weapons of World War II?*	2.8
Francium was discovered by Marguerite Perey in France (from which the element takes its name) in 1939.	*Where was Francium discovered by Marguerite Perey in 1939?*	3.8
Lazare Ponticelli was the longest-surviving officially recognized veteran... Although he attempted to remain with his French regiment, he eventually enlisted in...	*Did Lazare Ponticelli attempt to remain with his French regiment?*	4.4

Table 2: Example computer-generated questions, along with their mean ratings from Mechanical Turk.

Image Annotation for many purposes

Figure 1: Screenshot of the image annotation task.

- Goal: produce simple, full-sentence descriptions of images
- Motivations: scene understanding, generation of paraphrases, training an image labeler, ...

- Generate-Verify Setup:
  - ask for descriptions of 1000 images, 10 per HIT, $0.10 per HIT, 5 workers/HIT
  - judge for grammaticality/spelling without showing the picture: 5 per HIT (1 control), $0.08 per HIT, 3 workers/HIT
- Assessing the impact of a qualification test required to be able to work on the task:
  - grammar/spelling: detect whether there is an error
  - image content: choose the better description

Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J. (2010): Collecting Image Annotations Using Amazon's Mechanical Turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139–147, Los Angeles, CA, USA

Image Annotation Qualification Test for Grammar/Spelling

Are all of the words correctly spelled and correctly used?	Is the sentence grammatically correct?
A group of children playing with thier toys (N)	A man giving pose to camera. (N)
He accepts the crowd's praise graciously. (Y)	The white sheep walks on the grass. (Y)
The coffee is kept at a very hot temperture. (N)	She is good woman. (N)
A green car is parked in front of a resturant. (N)	He should have talk to him. (N)
An orange cat sleeping with a dog that is much larger then it. (N)	He has many wonderful toy. (N)
I ate a tasty desert after lunch. (N)	He sended the children home to their parents. (N)
A group of people getting ready for a surprise party. (Y)	The passage through the hills was narrow. (Y)
A small refrigerator filled with colorful fruits and vegetables. (Y)	A sleeping dog. (Y)
Two men fly by in a red plain. (N)	The questions on the test was difficult. (N)
A causal picture of a man and a woman. (N)	In Finland, we are used to live in a cold climate. (N)
Three men are going out for a special occasion. (Y)	Three white sheeps graze on the grassy field. (N)
Woman eatting lots of food. (N)	Between you and me, this is wrong. (Y)
Dyning room with chairs. (N)	They are living there during six months. (N)
A woman recieving a package. (N)	I was given lots of advices about buying new furnitures. (N)
This is a relatively uncommon occurance. (Y)	A horse being led back to it's stall. (N)

Table 3: The spelling and grammar portions of the qualification test. The test may be found on MTurk by searching for the qualification entitled "Image Annotation Qualification".

Image Annotation for many purposes II

Without qualification test
(1) lady with birds
(2) Some parrots are have speaking skill.
(3) A lady in their dining table with birds on her shoulder and head.
(4) Asian woman with two cockatiels, on shoulder
head, room with oak cabinets.,
(5) The lady loves the parrot
With qualification test
(1) A woman has a bird on her shoulder, and another bird on her head
(2) A woman with a bird on her head and a bird on her shoulder.
(3) A women sitting at a dining table with two small birds sitting on her.
(4) A young Asian woman sitting at a kitchen
table with a bird on her head and another on her shoulder.
(5) Two birds are perched on a woman sitting in a kitchen.

Figure 5: Comparison of captions written by Turkers with and without qualification test

- qualification test results in much higher worker quality: unqualified contained nonsensical responses and a lot of grammar errors
- verification not needed for qualified workers: simple pre-screening improves results a lot.

A video is worth 25 pictures per second...

http://www.cs.utexas.edu/users/ml/clamp/videoDescription/

- MSRVid corpus: same idea, but describing what can be seen in 2089 (short) videos
- this elicits descriptions of actions, rather than situations
- data was used in the SemEval tasks on Short Text Similarity from 2012
- Works in any language:

- A person is slicing a cucumber into pieces.
- A chef is slicing a vegetable.
- A person is slicing a cucumber.
- A woman is slicing vegetables.
- A woman is slicing a cucumber.
- A person is slicing cucumber with a knife.
- A person cuts up a piece of cucumber.
- A man is slicing cucumber.
- A man cutting zucchini.

Figure 1: Video and corresponding descriptions from MSRvid

English	85550	Hindi	6245	Romanian	3998	Slovene	3584
Serbian	3420	Tamil	2789	Dutch	2735	German	2326
Macedonian	1915	Spanish	1883	Gujarati	1437	Russian	1243
French	1226	Italian	953	Georgian	907	Polish	544

Chen, D. L. and Dolan, W.B. (2011): Collecting Highly Parallel Data for Paraphrase Evaluation. In the proceedings of The 49th Annual Meetings of the Association for Computational Linguistics (ACL), Portland, OR, USA

Let’s Crowdsource!
Find all the spelling and grammar errors

- VOTERS all over Europe have lost face in the EU because of its meddling in there lives, the EU Comission president said in Strassbourg. Public support have collapsed right across the EUs’ 28 member nations.
- In a astonishing confesion of failure, he added: “We are no longer respected in our countrys when we emphazise the need to give priority to the EU.”.
- His remarks were being seen as recognition of public revolsion at the EU ahead of Britains’ in-or-out referendum, on June 23 says the Express.

Hands up – how many errors?

Let's Crowdsource!
Find all the spelling and grammar errors

- VOTERS all over Europe have lost **faith** in the EU because of its meddling in **their** lives, the EU **Commission** president said in **Strasbourg**. Public support **has** collapsed right across the EU**'s** 28 member nations.
- In **an** astonishing **confession** of failure, he added: "We are no longer respected in our **countries** when we **emphasise** the need to give priority to the EU.".
- His remarks were being seen as recognition of public **revulsion** at the EU ahead of **Britain's** in-or-out referendum on June 23**,** says the Express.

13 Errors!

http://gibraltarpanorama.gi/15209/191118/a/eu-faces-ruin-voters-all-over-europe-lose-faith-in-the-eu-says-european-commissi

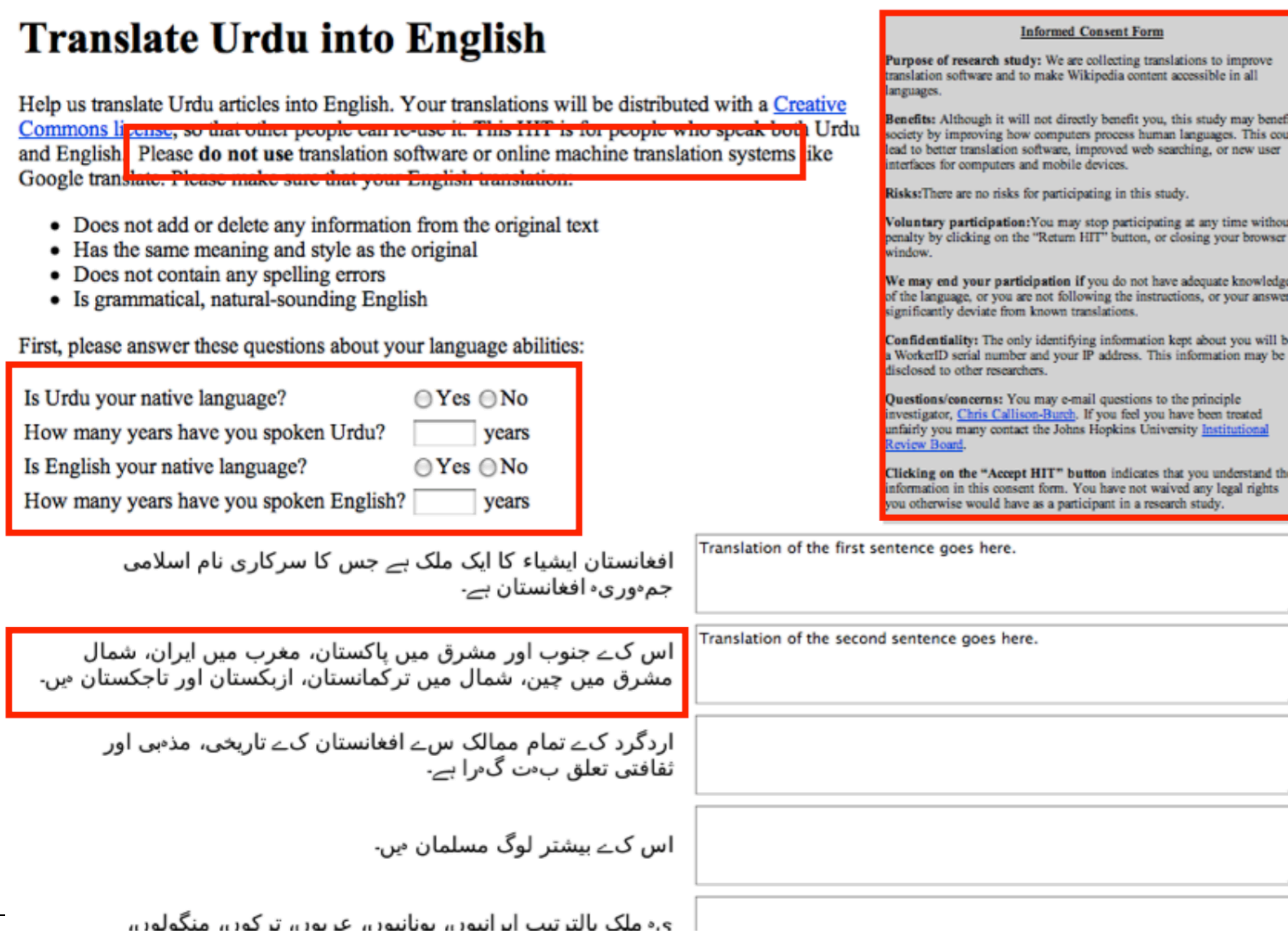
Crowdsourcing Translations
(materials from Chris Callison-Burch’s tutorial)

- Motivation:
  - Train a Machine Translation system
  - Existing parallel data does not cover all languages and domains
- Solution
  - use crowdsourcing for translation

- Zaidan&Callison-Burch’11 Setup on MTurk:
  - $0.10 to translate a sentence
  - $0.25 for post-editing 10 sentences
  - $0.06 to rank 4 translation groups

Zaidan, O. F. and Callison-Burch, C. (2011): Crowdsourcing translation: professional quality from non-professionals. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11), Vol. 1. pp. 1220-1229, Portland, OR, USA

Translation Interface on MTurk
(slide by Chris Callison-Burch – Task: translation into English)



Translate Urdu into English

Help us translate Urdu articles into English. Your translations will be distributed with a Creative Commons license, so that other people can re-use it. This HIT is for people who speak both Urdu and English. Please **do not use** translation software or online machine translation systems like Google translate. Please make sure that your English translation:

- Does not add or delete any information from the original text
- Has the same meaning and style as the original
- Does not contain any spelling errors
- Is grammatical, natural-sounding English

First, please answer these questions about your language abilities:

Is Urdu your native language?	◯ Yes ◯ No
How many years have you spoken Urdu?	years
Is English your native language?	◯ Yes ◯ No
How many years have you spoken English?	years

Informed Consent Form

Purpose of research study: We are collecting translations to improve translation software and to make Wikipedia content accessible in all languages.

Benefits: Although it will not directly benefit you, this study may benefi society by improving how computers process human languages. This cou lead to better translation software, improved web searching, or new user interfaces for computers and mobile devices.

Risks: There are no risks for participating in this study.

Voluntary participation: You may stop participating at any time withou penalty by clicking on the "Return HIT" button, or closing your browser window.

We may end your participation if you do not have adequate knowledg of the language, or you are not following the instructions, or your answer significantly deviate from known translations.

Confidentiality: The only identifying information kept about you will b a WorkerID serial number and your IP address. This information may be disclosed to other researchers.

Questions/concerns: You may e-mail questions to the principle investigator, Chris Callison-Burch. If you feel you have been treated unfairly you many contact the Johns Hopkins University Institutional Review Board.

Clicking on the "Accept HIT" button indicates that you understand th information in this consent form. You have not waived any legal rights you otherwise would have as a participant in a research study.

افغانستان ایشیاء کا ایک ملک ہے جس کا سرکاری نام اسلامی جمہوریہ افغانستان ہے۔	Translation of the first sentence goes here.
اس کے جنوب اور مشرق میں پاکستان، مغرب میں ایران، شمال مشرق میں چین، شمال میں ترکمانستان، ازبکستان اور تاجکستان ہیں۔	Translation of the second sentence goes here.
اردگرد کے تمام ممالک سے افغانستان کے تاریخی، مذہبی اور ثقافتی تعلق بہت گہرا ہے۔	
اس کے بیشتر لوگ مسلمان ہیں۔	
یہ ملک بالترتیب ایرانیوں، یونانیوں، عربوں، ترکوں، منگولوں	

Translation Verification Interface on MTurk
(slide by Chris Callison-Burch – Task: translation into English)

Vote for the best translation

Please read the sentences and vote on the one that you think is the best in each group. The sentences are translations that were produced by p who are not native English speakers. Their translations are often ungrammatical, misspelled, disfluent, or bad in other ways. Your goal is to t pick the best translation among the set. The one that you choose as the best will be forwarded on for editing, and it will undergo a variety of c quality control mechanisms before it is published.

You should consider the following factors when selecting one translation as the best:

- Does it make more sense than the others?
- Is the English reasonably good?
- Do the grammar and spelling require only minimal correction?

○	Experimentaions have proved that those rats on less calories diet have developed a tendency of not overcoming the flu virus .
○	in has been proven from experiments that rats put on diet with less calories had less ability to resist the Flu virus.
○	Experiments proved that mice on a lower calorie diet had comparatively less ability to fight the flu virus.
○	It was proved by experiments the low calories eaters mouses had low defending power for flue in ratio.

○	The research proved this old talk that decrease eating is useful in fever.
○	Research disproved the old axiom that " It is better to fast during fever"
○	research has proven this old myth wrong that its better to fast during fever.
○	This Research has proved the very old saying wrong that it is good to starve while in fever.

○	According to the scientist a patient should eat more while in fever.
○	According to scientists, eat a lot during fever.
○	Eat and drink more in fever according to scientists.
○	according to the scientists one should eat a lot during fever.

Quality Control Model for Translation
(slide by Chris Callison-Burch – Task: translation into English)

- Sentence features
  - Language model probability
  - Ratio of source / target sentence lengths
  - Web n-gram match percentage
  - Translation edit rate to other translators
- Worker features
  - Aggregate of sentence feature scores
  - Self-reported language abilities (Is native speaker? How long speaking?)
  - Worker location (Pakistan? India?)
- Ranking features (based on second pass vote)
- Calibration feature (Bleu against professionals)

BLEU Translation Evaluation Metric

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

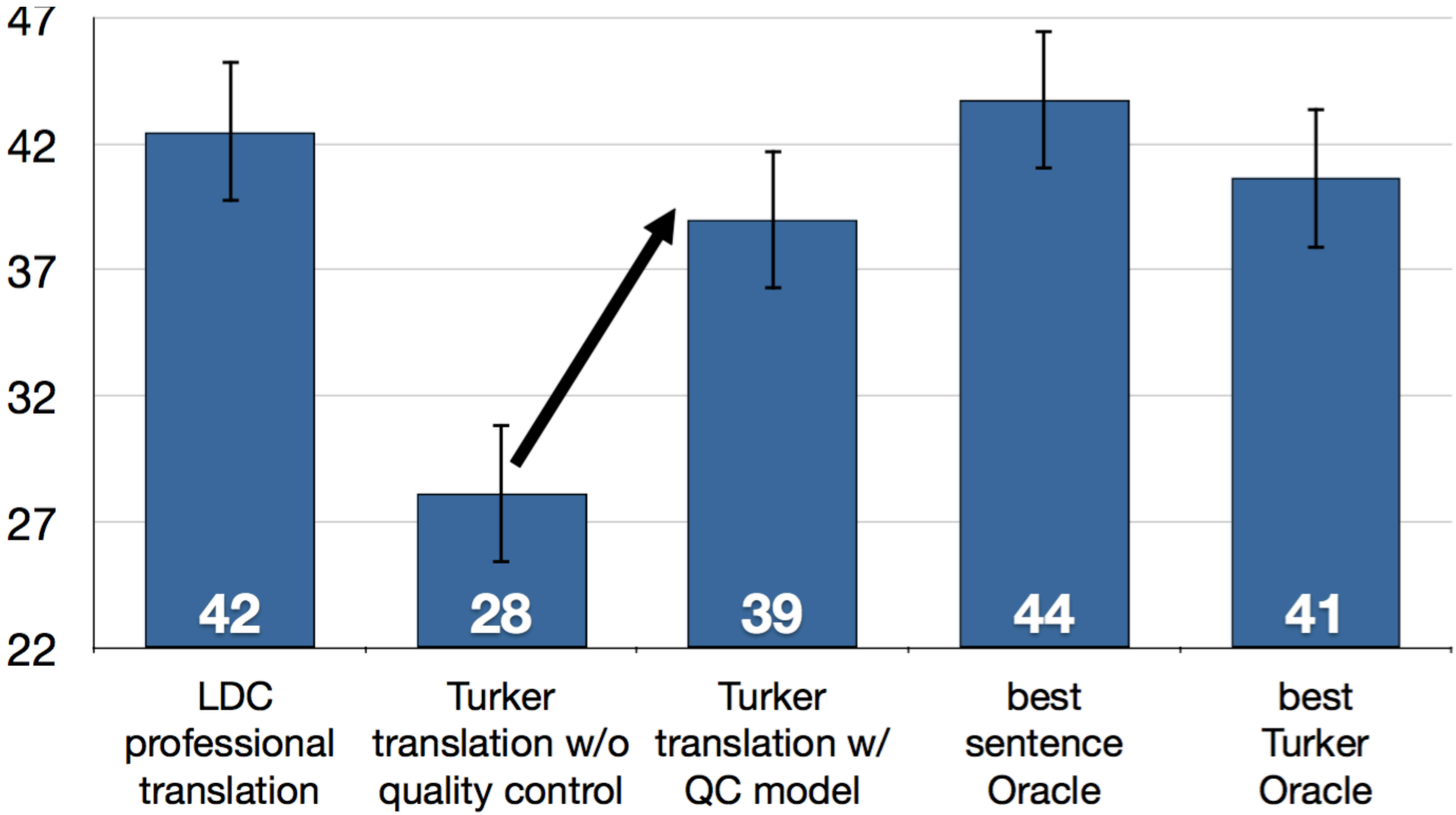
BLEU4 formula

(counts n-grams up to length 4)

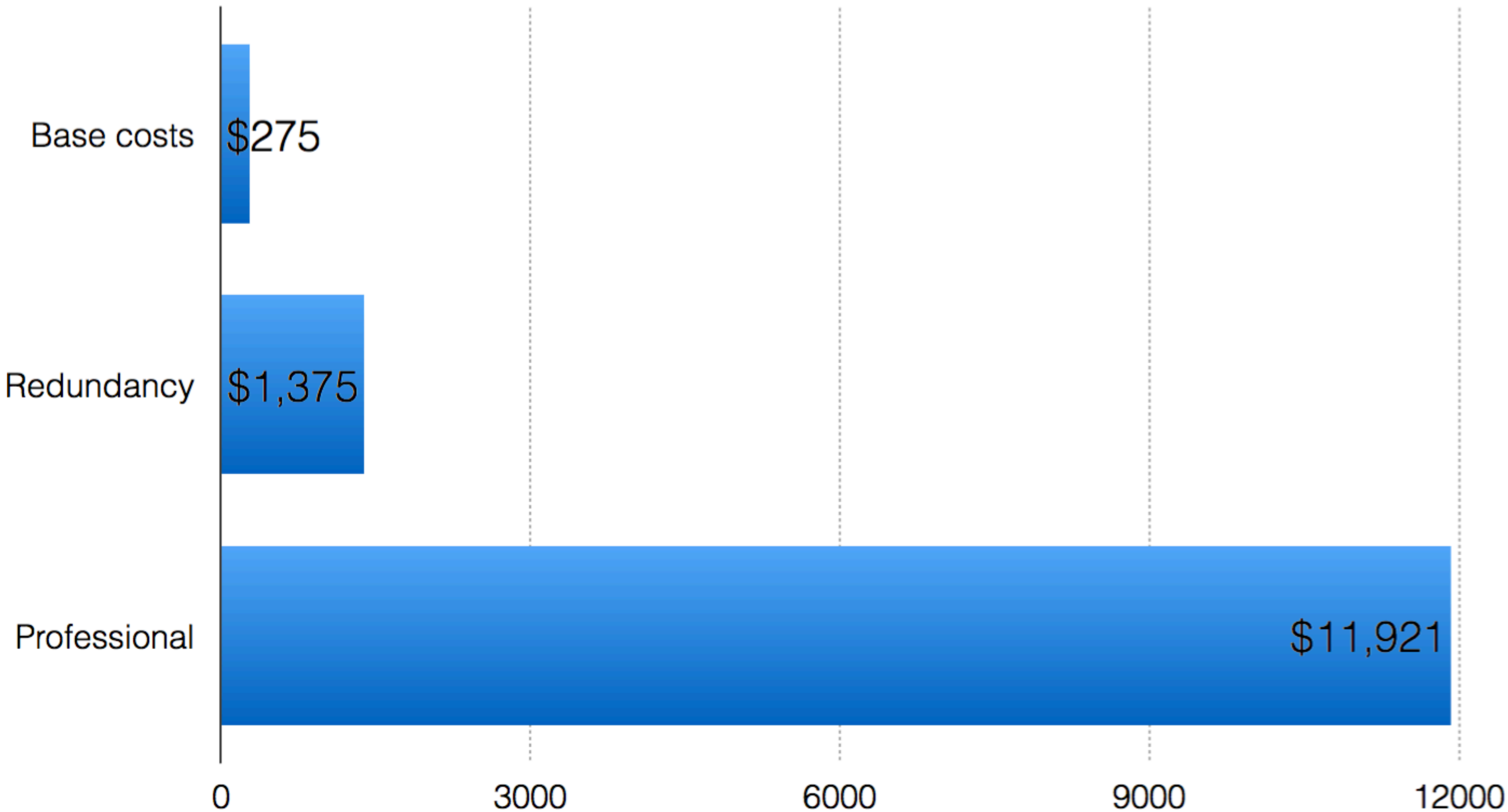
exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0))

p1 = 1-gram precision
p2 = 2-gram precision
p3 = 3-gram precision
p4 = 4-gram precision

Crowds approaching professional quality
(slide by Chris Callison-Burch – Task: translation into English)



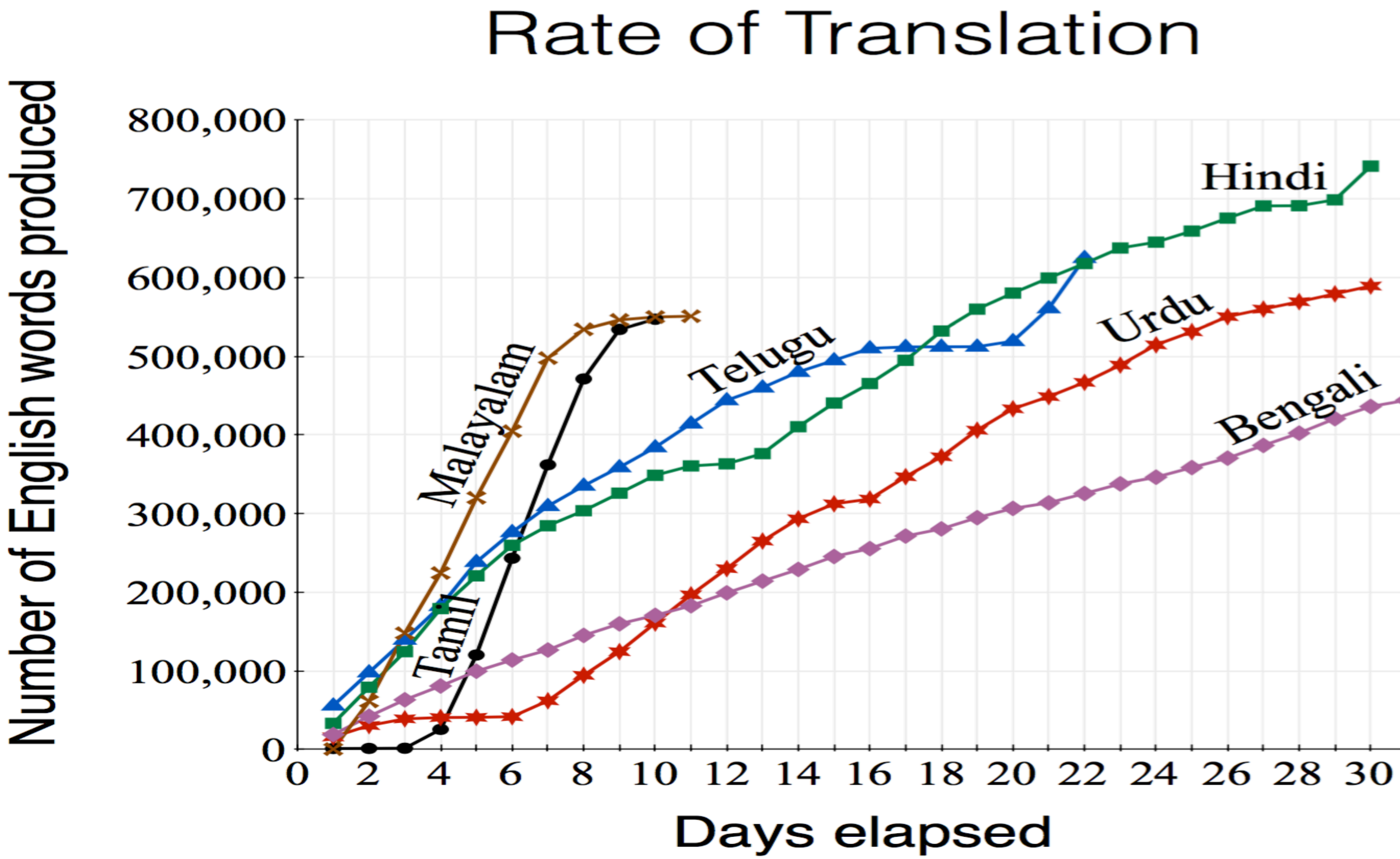
Crowds not approaching professional's costs
(slide by Chris Callison-Burch – Task: translation into English)



Translator Availability on MTurk
(slide by Chris Callison-Burch – Task: translation into English)

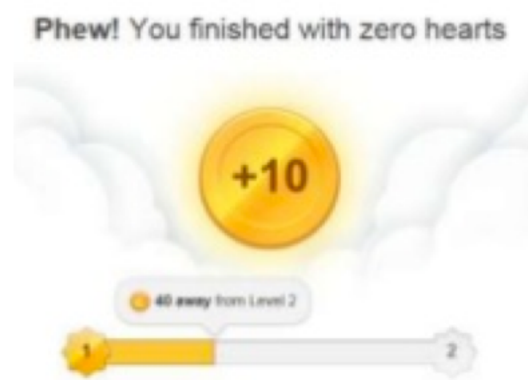
workers	quality	speed	
many	high	fast	Dutch, French, German, Gujarati, Italian, Portuguese, Romanian, Serbian, Spanish, Tagalog, Telugu
		slow	Arabic, Hebrew, Irish, Punjabi, Swedish, Turkish
	medium or low	fast	Hindi, Marathi, Tamil, Urdu
		slow	Bengali, Bishnupriya Manipuri, Cebuano, Chinese, Nepali, Newar, Polish, Russian, Sindhi, Tibetan
few	high	fast	Bosnia, Croatian, Macedonian, Malay, Serbo-Croatian
		slow	Afrikaans, Albanian, Aragonese, Asturian, Basque, Belarusian, Bulgarian, Central Bicolano, Czech, Danish, Finnish, Galacian, Greek, Haitian, Hungarian, Icelandic, Ilokano, Indonesian, Japanese, Javanese, Kapampangan, Kazakh, Korean, Lithuanian, Low Saxon, Malagasy, Nor- wegian (Bokmal), Sicilian, Slovak, Slovenian, Thai, Ukranian, Uzbek, Waray-Waray, West Frisian, Yoruba
	medium or low	slow	Amharic, Armenian, Azerbaijani, Breton, Catalan, Georgian, Latvian, Luxembourgish, Neapolitian, Norwegian (Nynorsk), Pashto, Piedmontese, Somali, Sudanese, Swahili, Tatar, Vietnamese, Walloon, Welsh
none			Esperanto, Ido, Kurdish, Persian, Quechua, Wolof, Zazaki

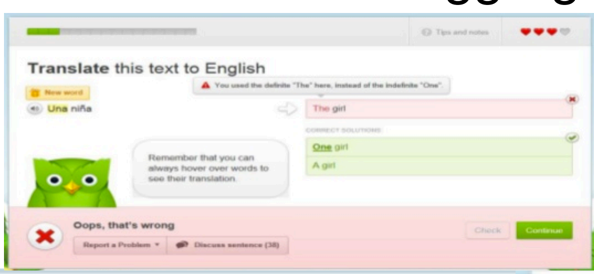
Translation Speed of MTurk for different languages
(slide by Chris Callison-Burch – Task: translation into English)

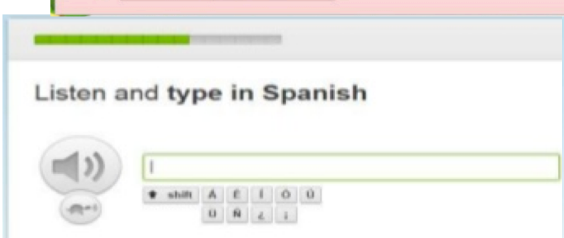


Duolingo Commercial Model

- Incentive for crowdworkers: learn a language!
- Founder: Luis von Ahn, see also: ESP game, reCaptcha
- Language Learners translate sentences according to their level.
- More advanced learners correct these.
- also: collection of speech corpora
- Translations are aggregated and sold as a service



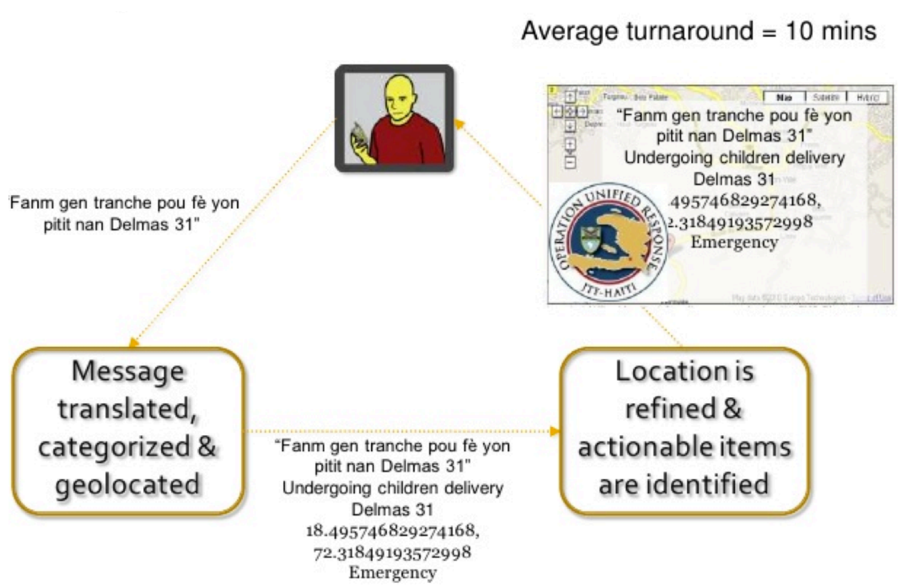




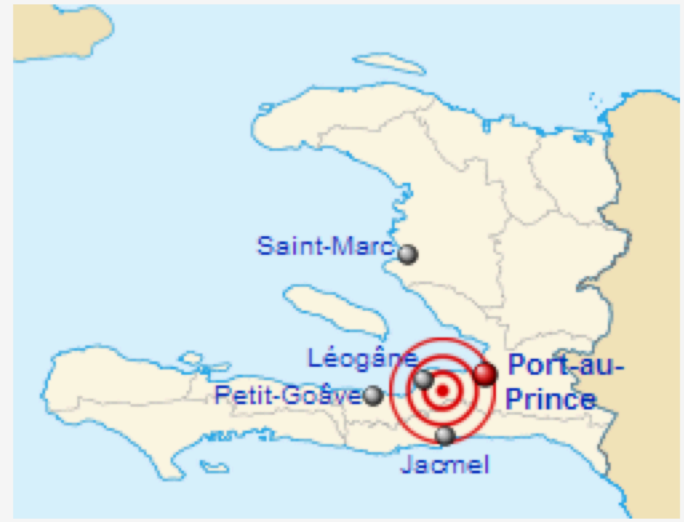
http://www.slideshare.net/katfish2008/duolingo-powerpoint?qid=4c2767b8-9f8a-4381-98e8-2251eb364560&v=&b=&from_search=1

Crowdsourced Translation for Emergency Response

- 2010 Haiti Earthquake
- Text messaging is the only popular and working communication channel
- Aid personnel does not speak Creole
- "Mission 4636" launched in under 2 days, both volunteer and paid crowdwork



2010 Haiti earthquake



Date	12 January 2010
Origin time	16:53
Magnitude	7.0 M_w
Depth	13 km (8.1 mi)
Epicenter	18°27′25″N 72°31′59″W
Areas affected	Haiti, Dominican Republic
Max. intensity	MM X[1] (Extreme)
Peak acceleration	0.5 *g*[2]
Tsunami	Yes (localized)[3]
Casualties	100,000 to 316,000 deaths (the

http://www.slideshare.net/wwrob/realtime-crowdsourced-translation-for-emergency-response-and-beyond?qid=7d7e3002-ebe4-4afb-9872-be1b0c45edd7&v=&b=&from_search=1

In a Nutshell: Learned in Lesson 4

- Many sample projects for NLP tasks
- Introduction to many NLP problems
- Different quality control mechanisms in practice