A Framework for Conducting Non-Expert Translations and Summarizations

Christopher Harris

Department of Computer Science

SUNY Oswego

Photo by CEphoto, Uwe Aranas, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=34012183

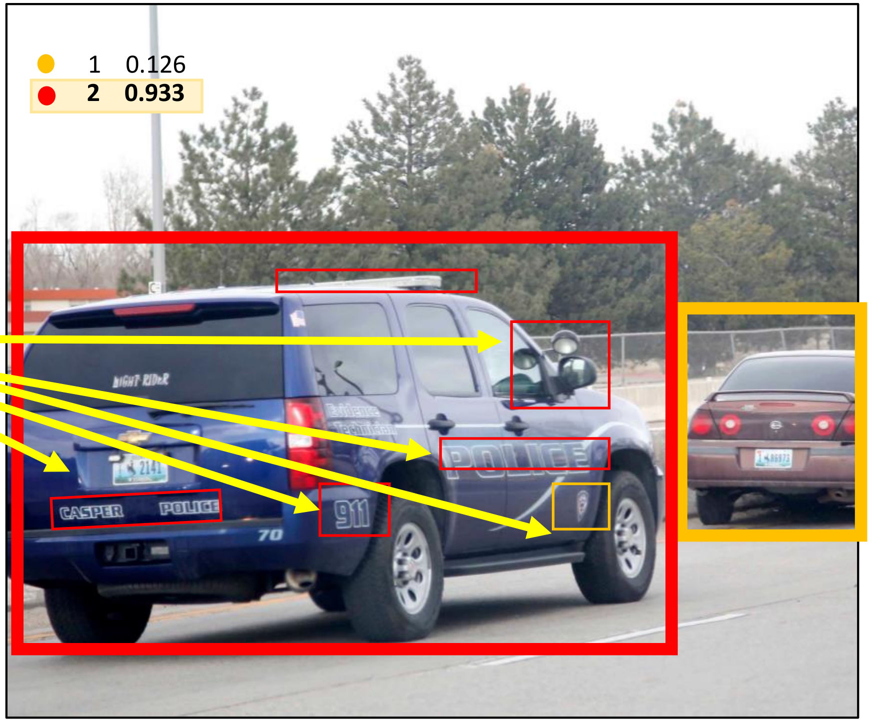
About me

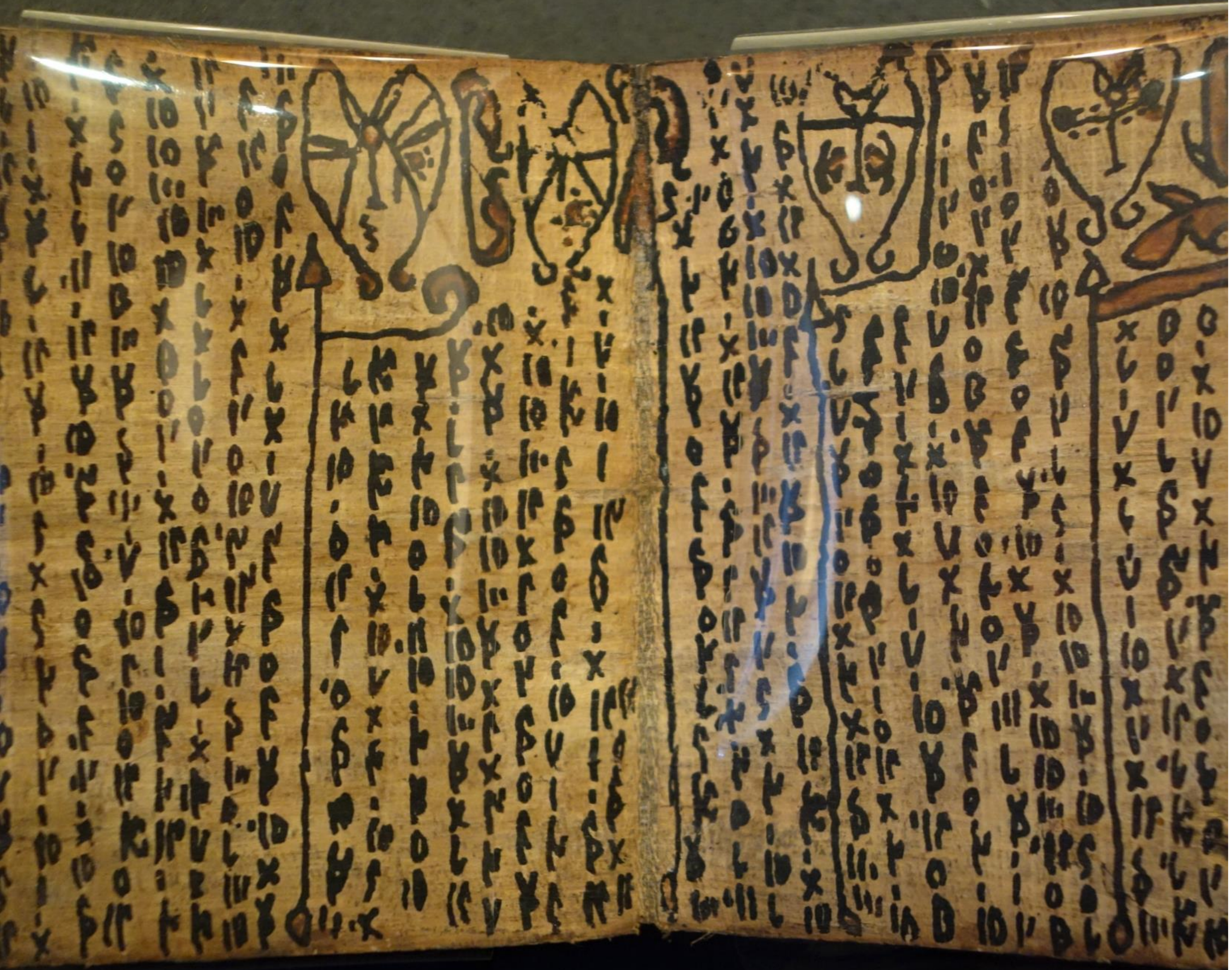
- Assistant Professor at SUNY Oswego
- Focus on Human Computation
- Research evaluates tradeoffs in using humans and computers for a variety of tasks
    - decision-making
    - knowledge creation
    - games and incentives

Other current research

- Using human computation (HC) in police car identification
- HC trains machine learning (ML) algorithms
- ML algorithms power augmented reality
- Real time decisions

The crowd identifies these features and a probability is assigned. This trains an ML algorithm



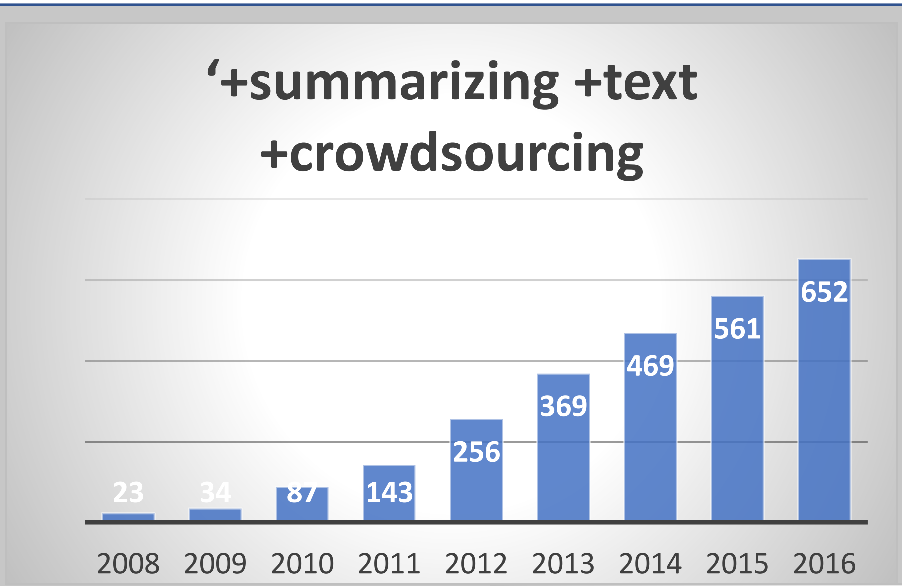


By Daderot - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=31742989

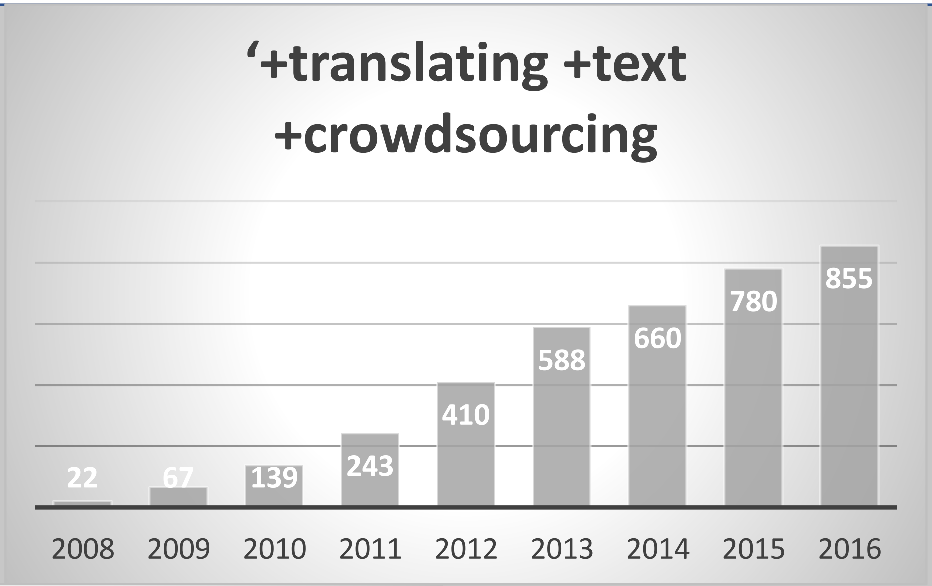
Focus on Crowdsourcing methods

- Applying them to NLP
  - Rare texts
  - Low-resource languages
- Text Summarizations
  - Children
  - Elderly
- Transcriptions (?)

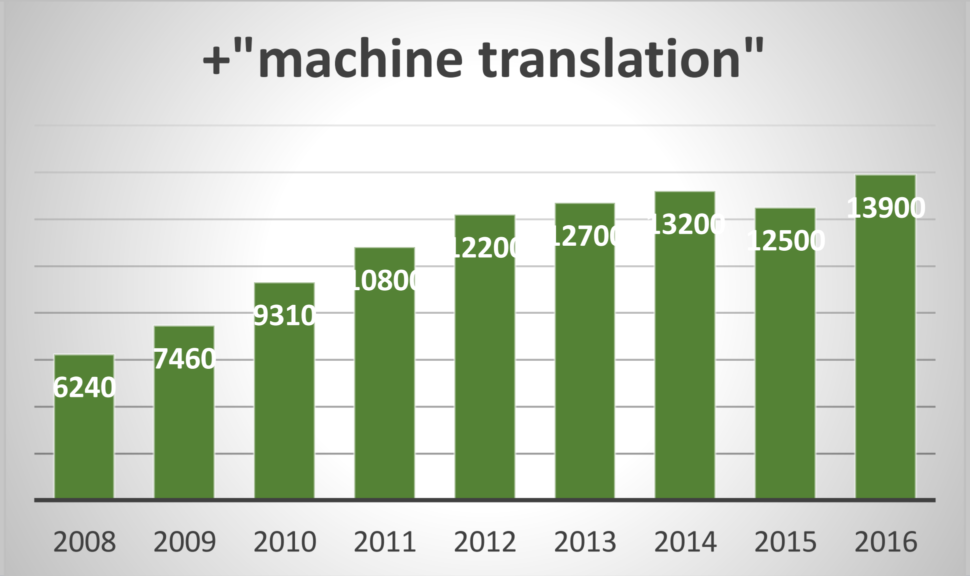
Use of the Crowd for NLP tasks



‘+summarizing +text +crowdsourcing



‘+translating +text +crowdsourcing

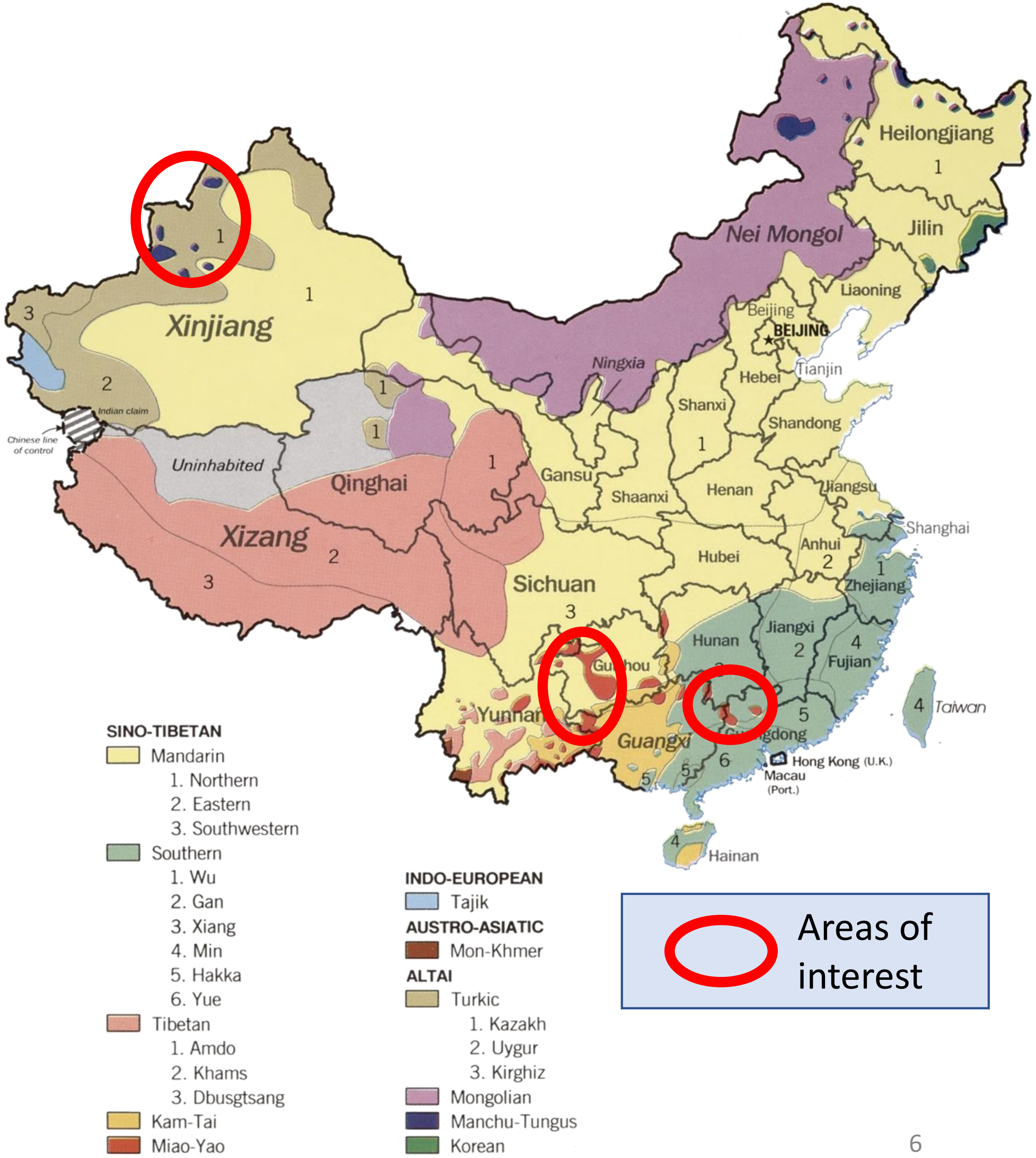


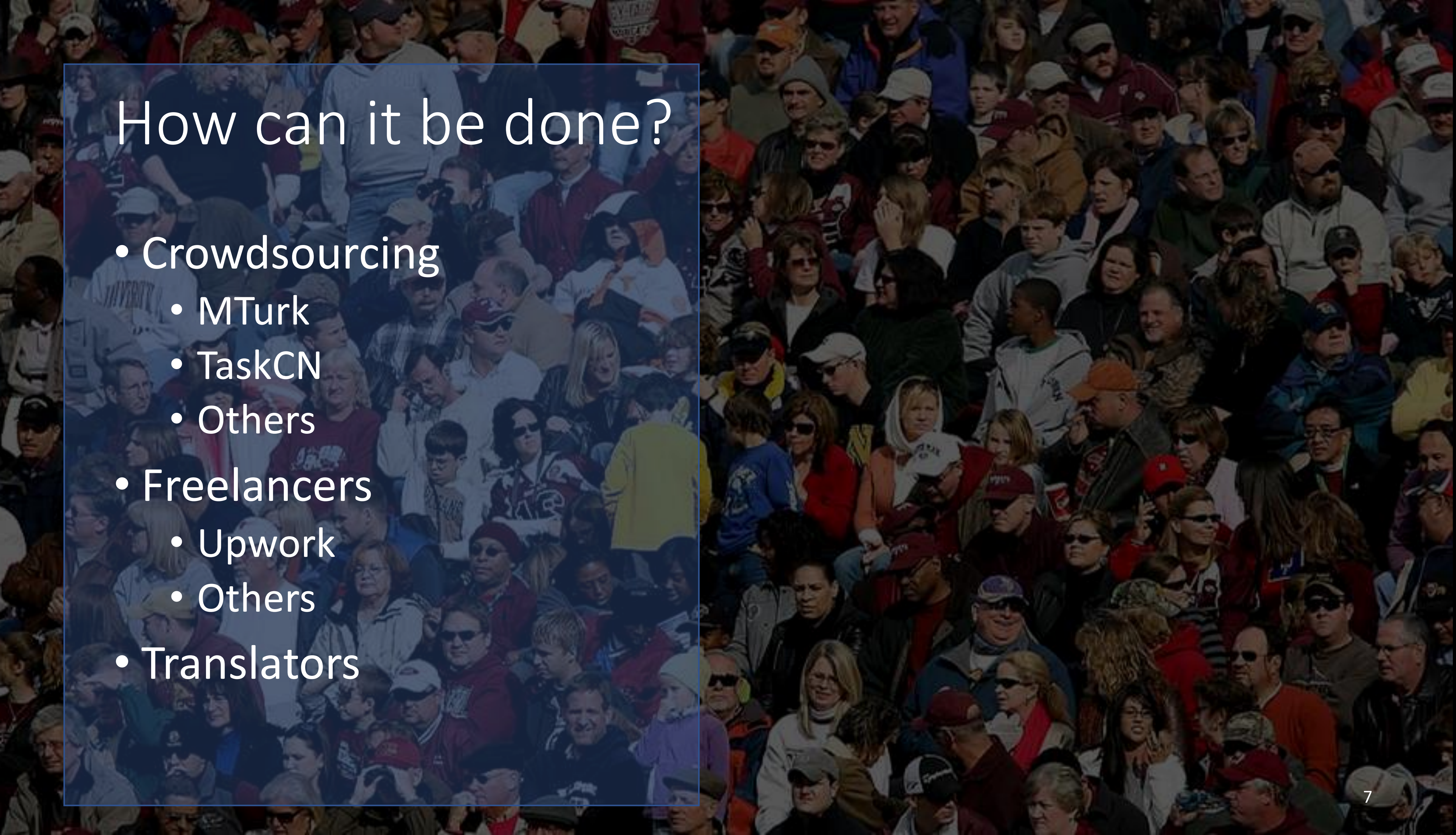
+"machine translation"

Google scholar
Conducted on 4/7/17

Linguistic groups of China

Public Domain, https://commons.wikimedia.org/w/index.php?curid=1466708





How can it be done?

- Crowdsourcing
  - MTurk
  - TaskCN
  - Others
- Freelancers
  - Upwork
  - Others
- Translators

How else can it be done?

- What about Edu-sourcing?
- Using students (high school and above) to perform translations and text summarizations

Objectives of this talk…

1. Describe

a framework for crowdsourcing both translations and text summarizations

2. Examine

some recent empirical experiments conducted using this framework.

3. Evaluate

some design elements, including

- the number (depth) of crowdworkers needed for different tasks in the framework
- how this depth affects output quality and task completion time.

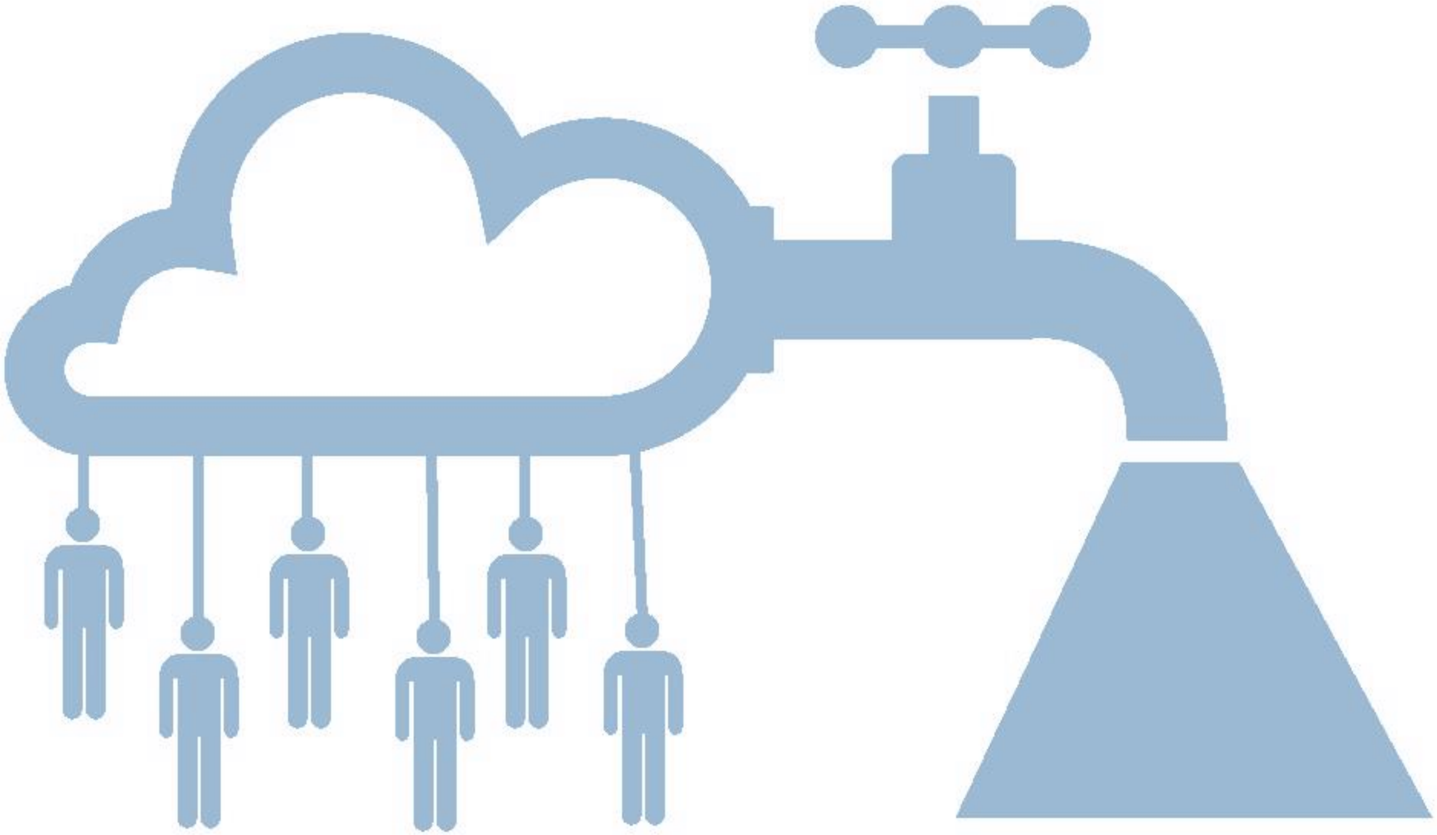
Framework

Translations using the Crowd

A well-trodden path

- **Snow et. al. (2008)**
  - One of the first to use Mturk for translations
  - Used Majority voting
- **Callison-Burch (2009)**
  - Used crowd output to score MT translations
- **Zaidan & Callison-Burch (2011)**
  - Split up document into snippets
  - Redundancy (parallel tasks) built in

Translations using the Crowd

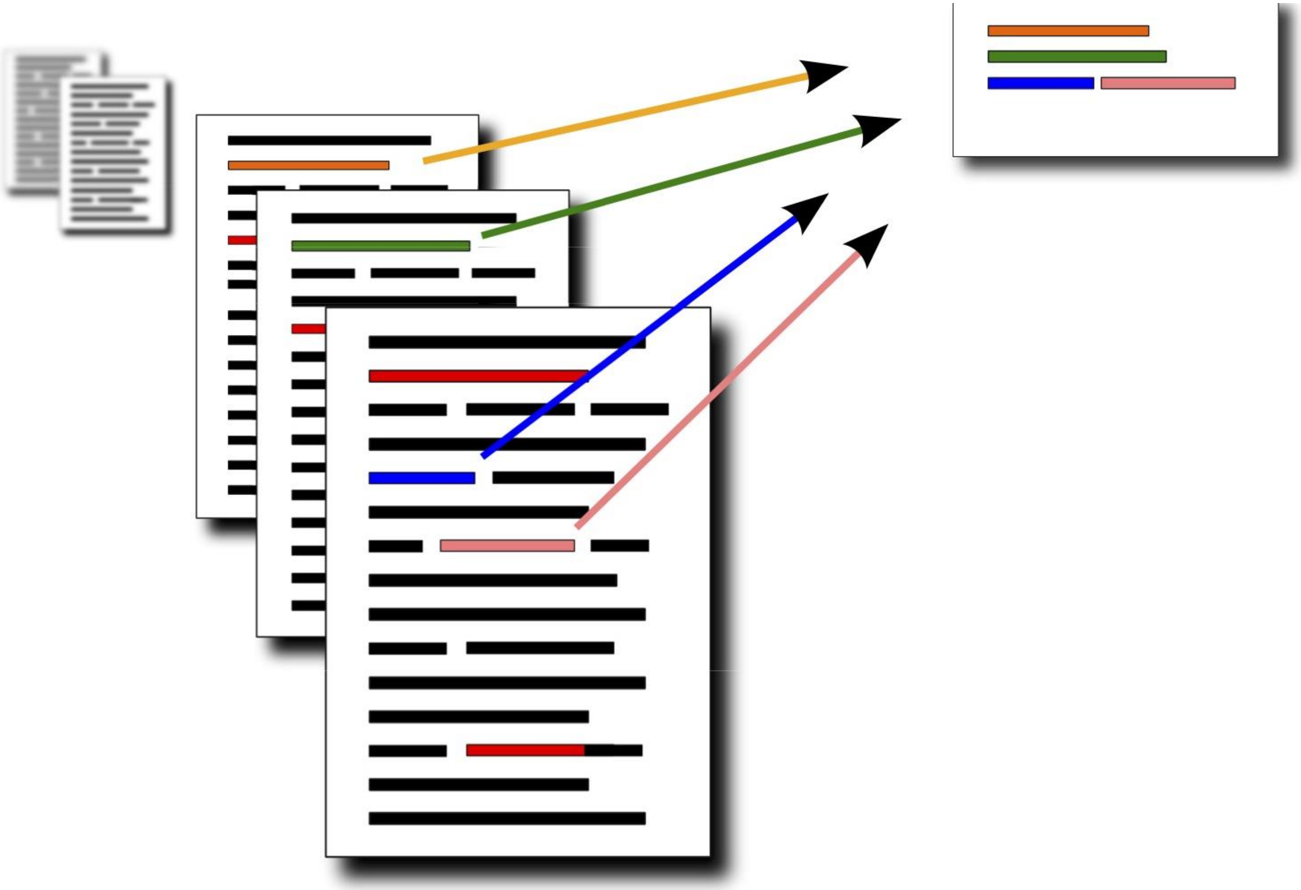


- **Ambati et. al. (2012)**
  - Annotations from multiple turkers
  - Examined quality vs. cost
- **Yan (2014)**
  - Two-step approach introduced
    - Translator
    - Editor
  - Relationship between the two improves reliability

Text Summarizations using the Crowd

Fewer Examples of Empirical Work

- Hourcade and Gehrt, (2015)
- used crowdworkers in a two-step process:
    - first to summarize ACOVE medication warnings
    - vote for the best summarization



Text Summarizations using the Crowd

El-Haj et al. (2010)

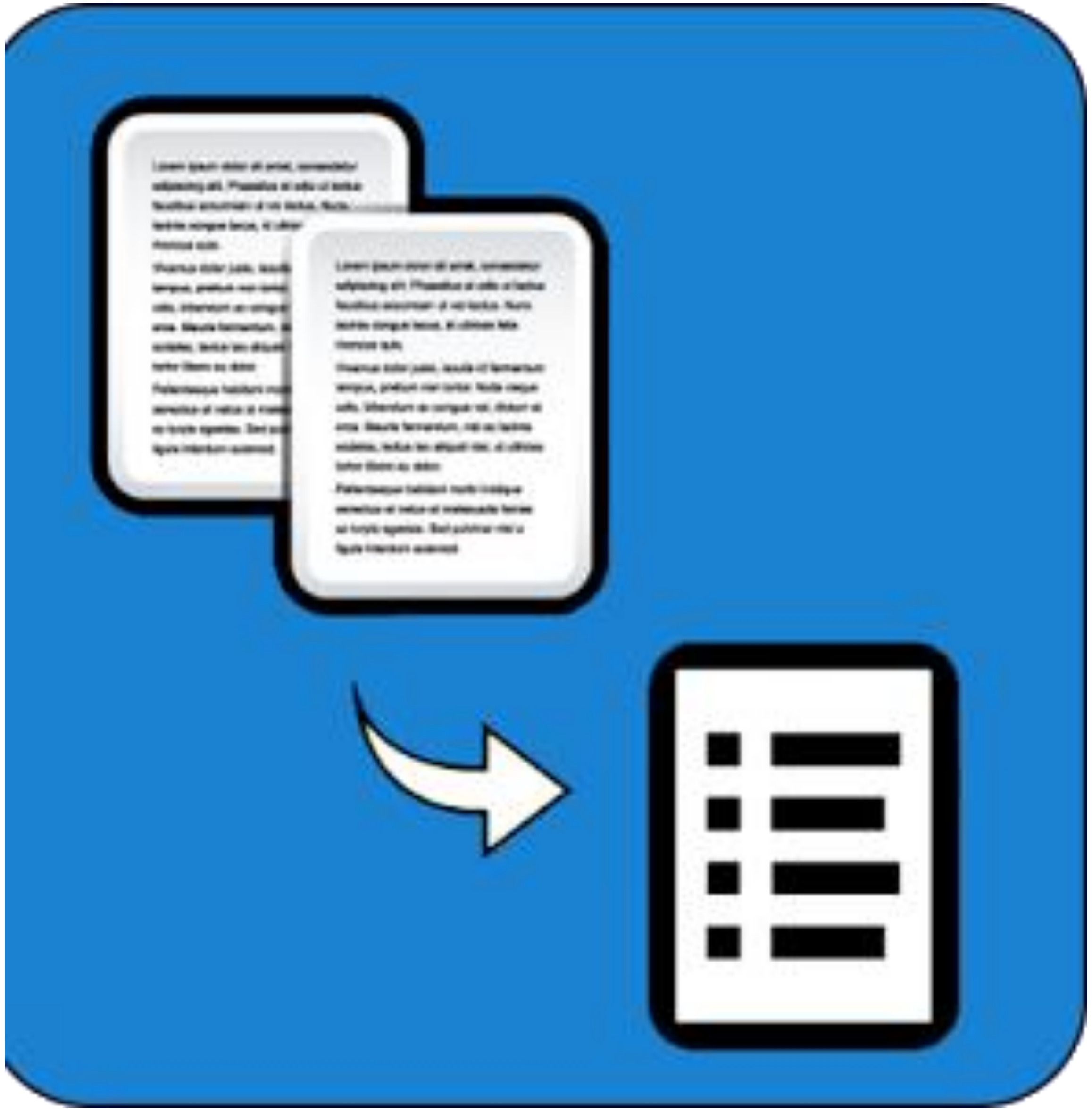
- used AMT to collect a corpus of single-document summaries from Wikipedia and newspaper articles in Arabic.
- Produced by extracting the most relevant sentences of the source document.

Text Summarizations using the Crowd

Buzek et al.(2010)

Mturk used to create paraphrase lattices as MT inputs.

- create the paraphrase lattices
- verify the generated paraphrases



Missing a bigger picture?

- Hard to say a technique works without considering the entire model!
- For example, consider a model:
  1. Divide document into snippets
  2. Translate
  3. Recombine snippets into document

  …But did the recombined document lose context and flow?
- Using one translator vs. many translators

We seek a framework with the following qualities

1

Robust:

- Our framework should be impervious to low-quality inputs from a malicious crowdworker.

2

Verifiable:

- Should be able to perform an evaluation of outputs after each crowdworker-dependent step in our framework.

3

Consistent:

- The same inputs should produce approximately the same outputs, even with different crowdworkers.

4

Flexible:

- As few components as possible should rely exclusively on multi- and bilingual crowdworkers.

Benefits of a Framework in CS/NLP

- Reproducible and repeatable
- Permits critical evaluation of assumptions
- Focus on the components can be done iteratively
- Constant improvement through refinements

Need a framework

Crowdsourcing-dependent components to include in the framework….

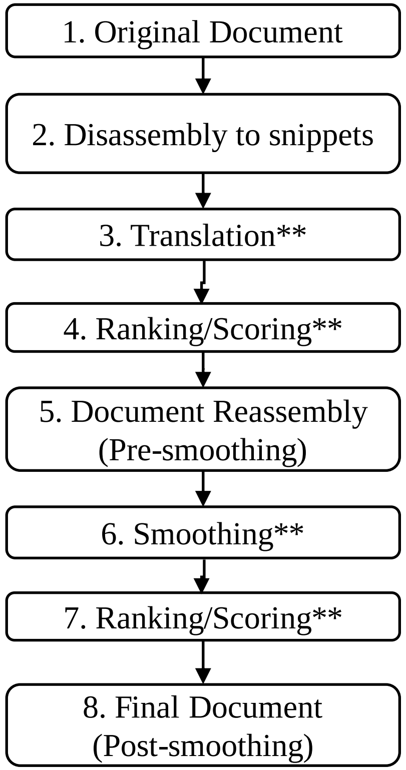
- **Ranking:**
  - Also called voting
  - Asks crowdworkers to place text in order of relative preference.
  - Helpful in situations where users have few choices and can clearly discriminate between the choices.
  - Can use a single-winner technique (e.g., Borda counting) or a multi-round technique
- **Scoring:**
  - Also called rating
  - Asks crowdworkers to provide a score to each text on a Likert scale.
  - Preferable when there are too many choices available to the worker to determine a clear relative preference.

Other components to include in the framework….

- **Translation/Summarization:**
  - Core essential component
  - Translated/summarized versions of the input text are generated
- **Disassembly/Reassembly:**
  - Divide a document (or set of documents) into snippets
  - Recombine the translated/summarized segments into a single document
  - Usually done through automation

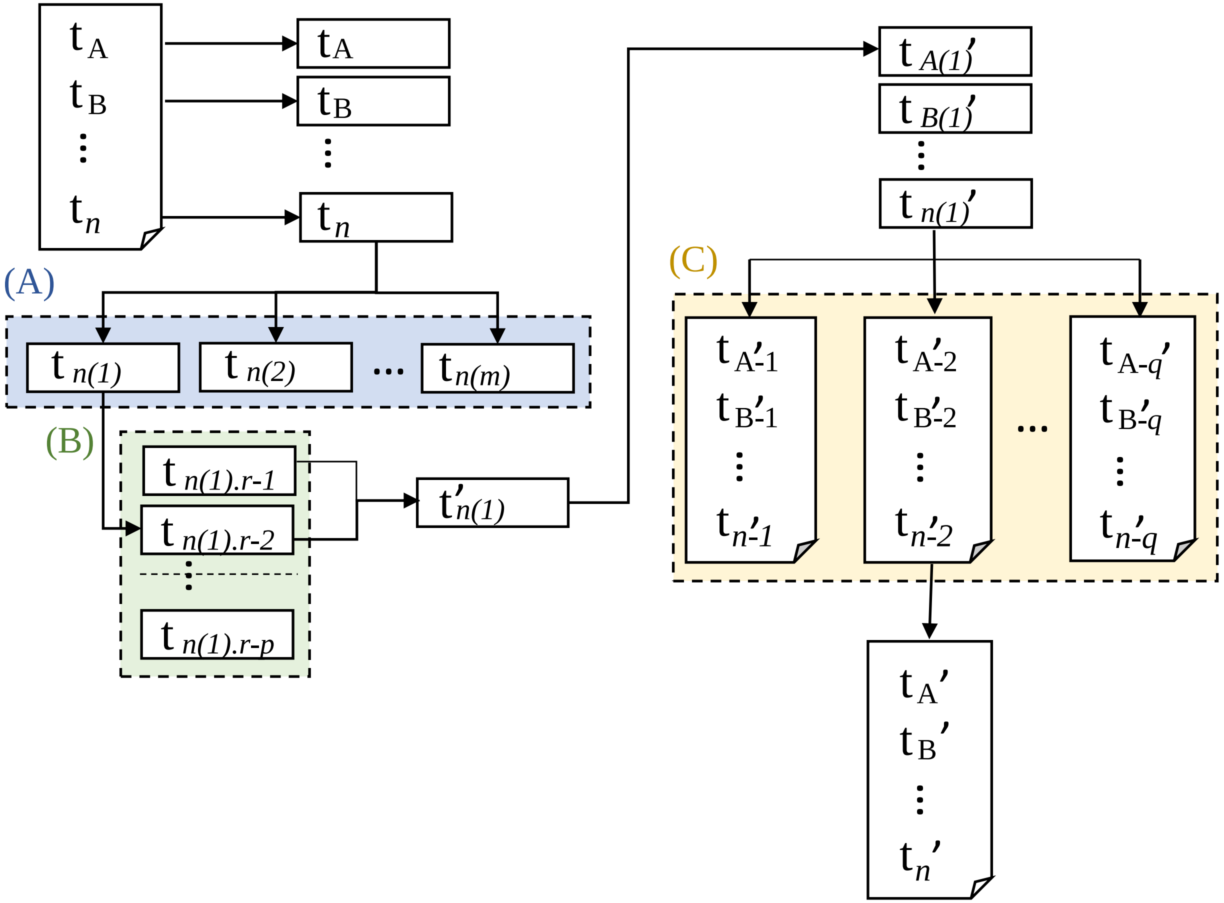
Introducing a Framework

Conceptual View



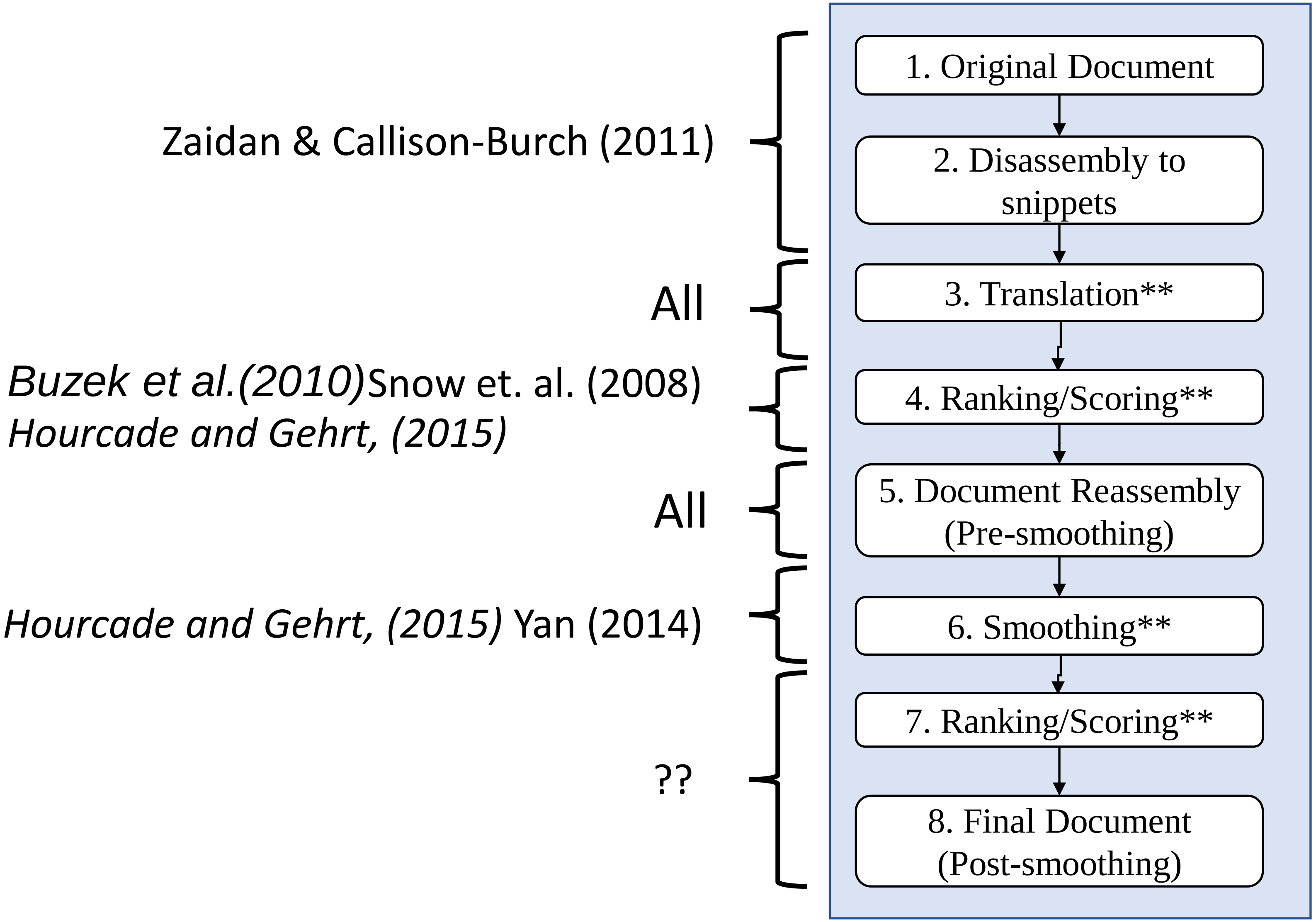
1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

Componentized View


t_A t_B ⋮ t_n
t_A
t_B
⋮
t_n
(A)
$t_{n(1)}$ $t_{n(2)}$ … $t_{n(m)}$
(B)
$t_{n(1).r\text{-}1}$
$t_{n(1).r\text{-}2}$
⋮
$t_{n(1).r\text{-}p}$
$t'_{n(1)}$
$t_{A(1)}'$
$t_{B(1)}'$
⋮
$t_{n(1)}'$
(C)
$t_{A'\text{-}1}$ $t_{B'\text{-}1}$ ⋮ $t_{n'\text{-}1}$
$t_{A'\text{-}2}$ $t_{B'\text{-}2}$ ⋮ $t_{n'\text{-}2}$
…
$t_{A\text{-}q'}$ $t_{B'\text{-}q}$ ⋮ $t_{n'\text{-}q}$
t_A' t_B' ⋮ t_n'

** = crowd-assisted

Framework Elements Previously Explored in CS/NLP



Zaidan & Callison-Burch (2011)

1. Original Document

2. Disassembly to snippets

All

3. Translation**

*Buzek et al.(2010)*Snow et. al. (2008)
Hourcade and Gehrt, (2015)

4. Ranking/Scoring**

All

5. Document Reassembly (Pre-smoothing)

Hourcade and Gehrt, (2015) Yan (2014)

6. Smoothing**

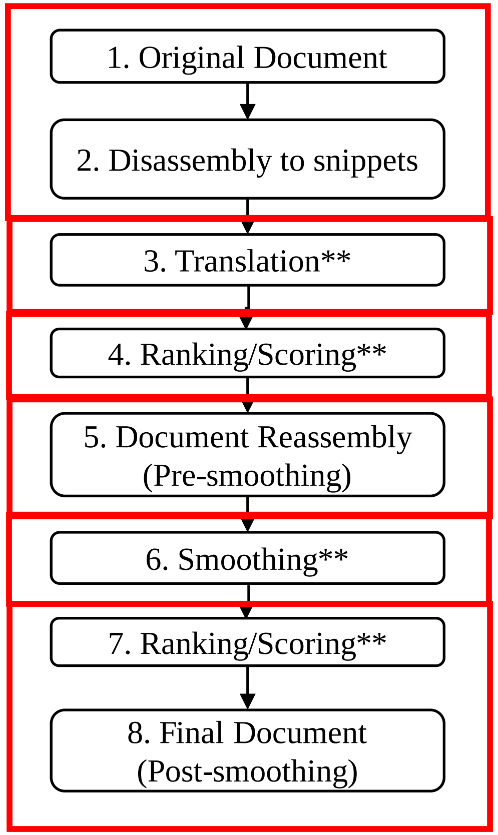
??

7. Ranking/Scoring**

8. Final Document (Post-smoothing)

** = crowd-assisted

Framework Components



1. Original Document

2. Disassembly to snippets

3. Translation**

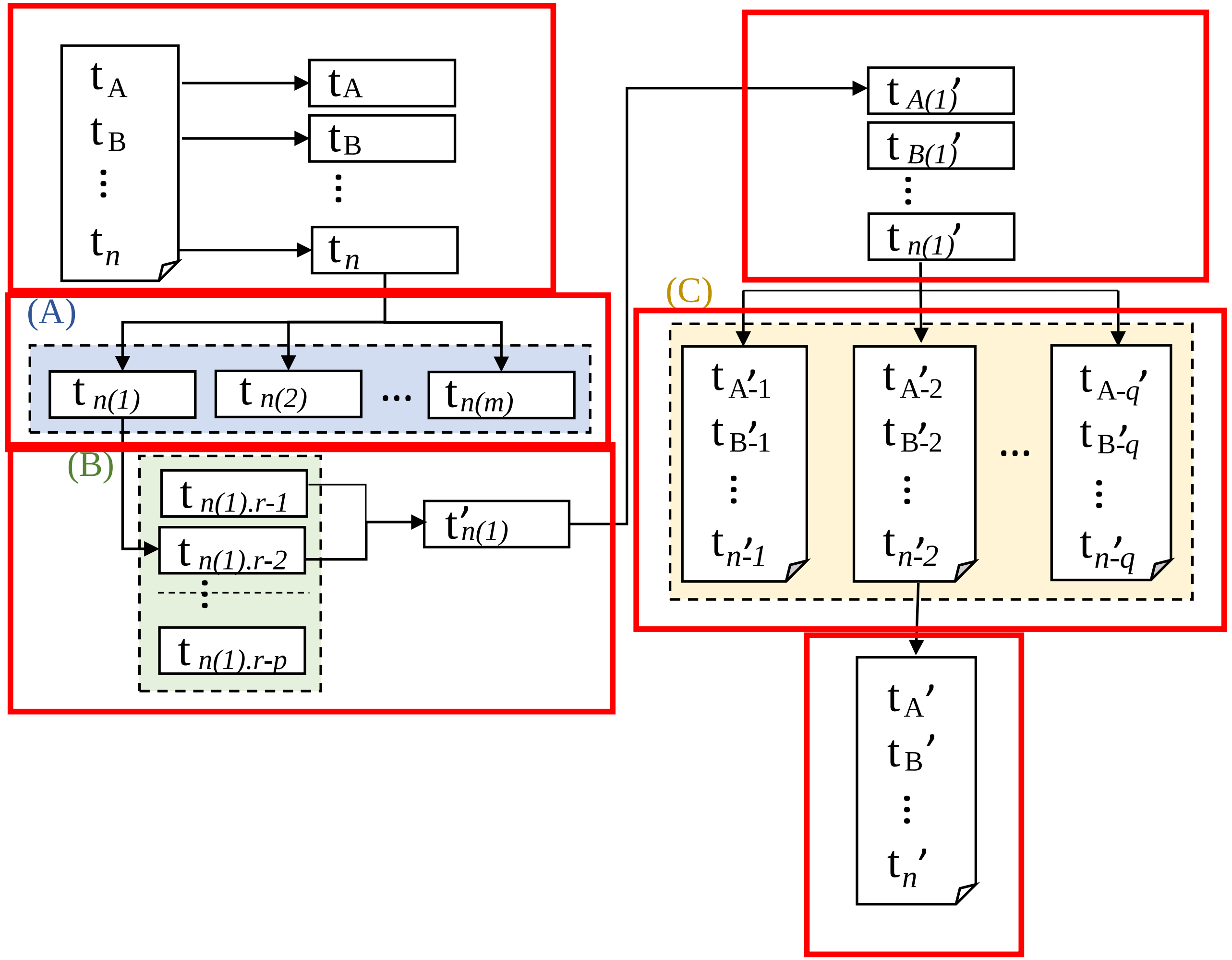
4. Ranking/Scoring**

5. Document Reassembly (Pre-smoothing)

6. Smoothing**

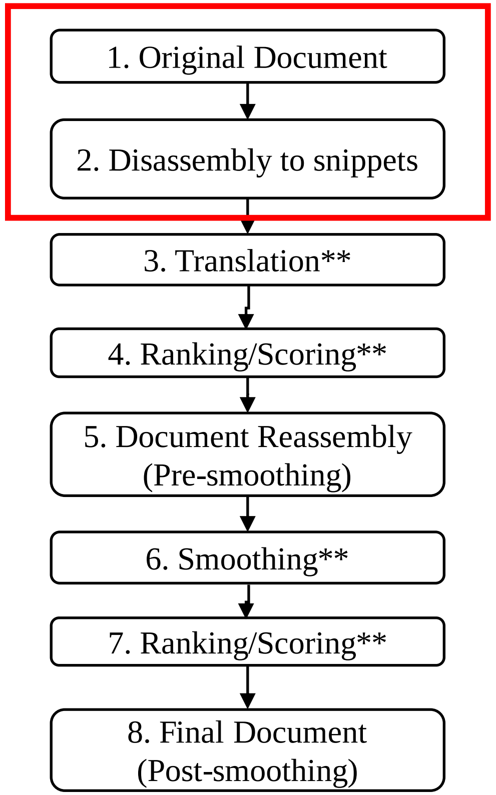
7. Ranking/Scoring**

8. Final Document (Post-smoothing)

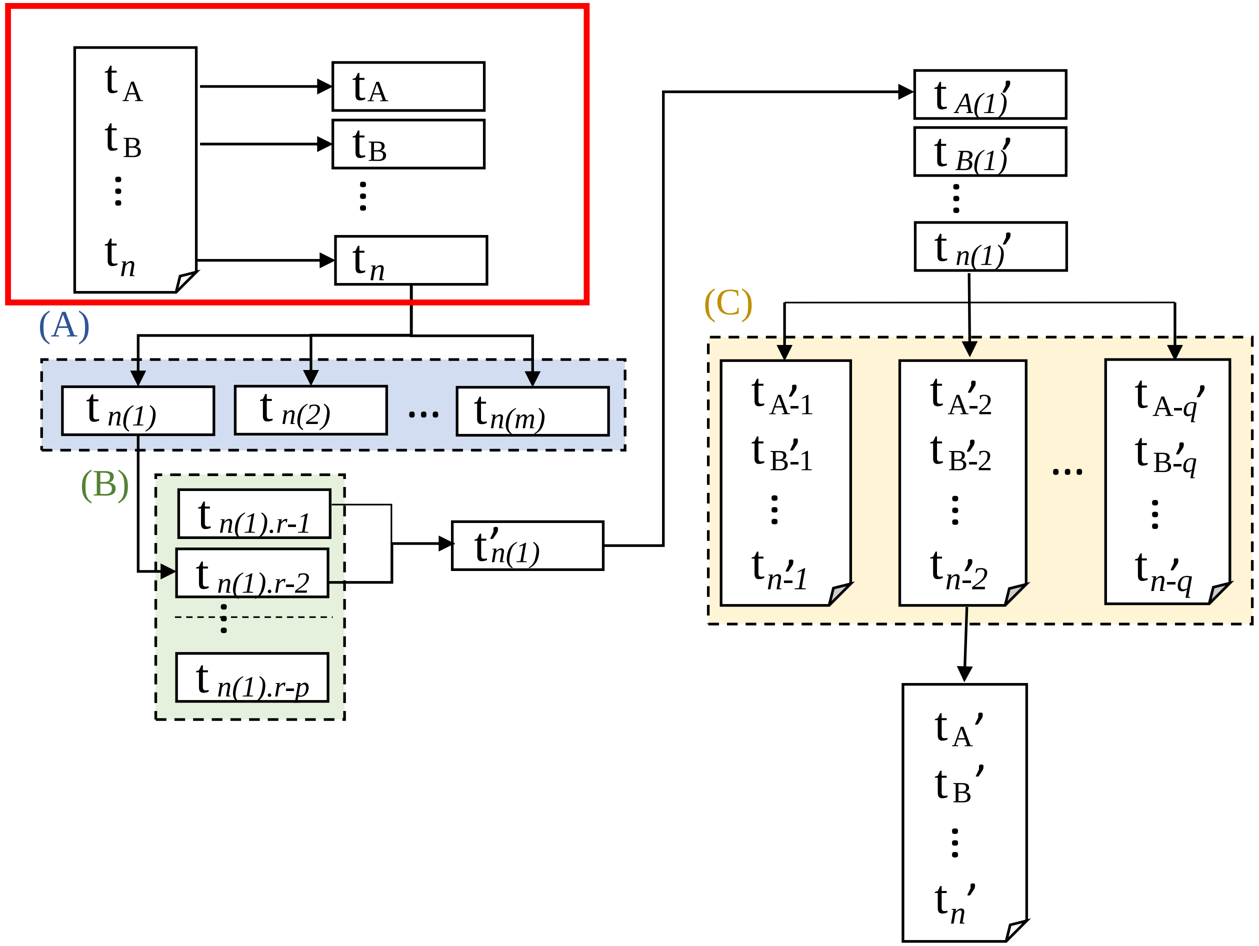


** = crowd-assisted

Framework: Disassembly

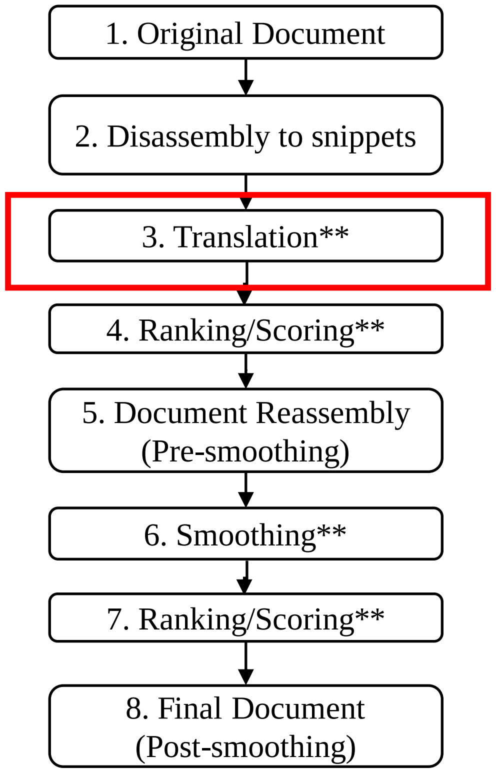


1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)



** = crowd-assisted

Framework: Translation



1. Original Document

2. Disassembly to snippets

3. Translation**

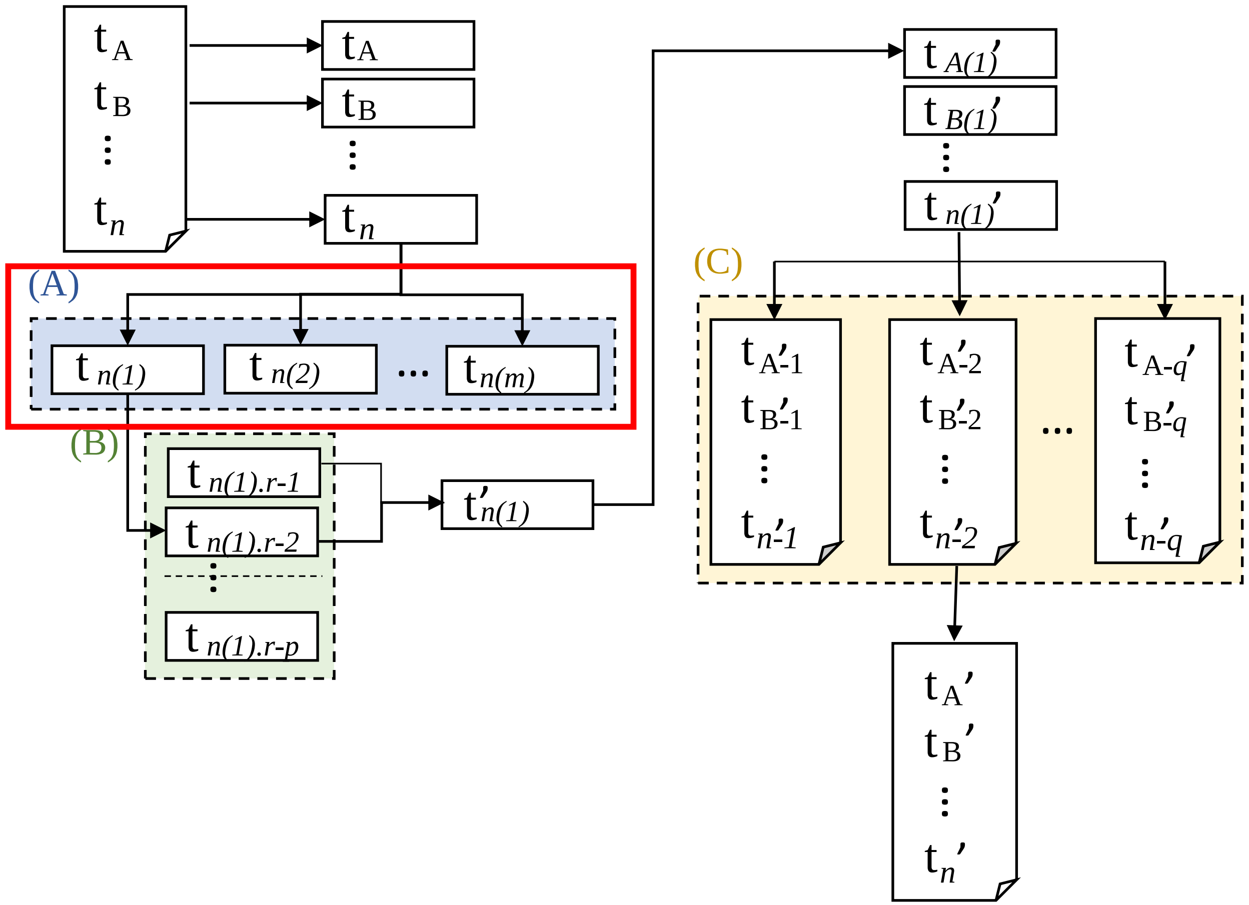
4. Ranking/Scoring**

5. Document Reassembly (Pre-smoothing)

6. Smoothing**

7. Ranking/Scoring**

8. Final Document (Post-smoothing)



** = crowd-assisted

Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

From Google Translate:

初步研究表明，一旦孩子承认自己的父母，他们常常被怀疑

Preliminary studies have shown that once children recognize their parents, they are often suspected

Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

初步研究表明，一旦一个孩子向父母坦白，他们往往会被怀疑*

Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical*

初步研究显示，一旦孩子向父母坦白，他们往往难以置信

Preliminary studies have shown that once children are confessed to their parents, they are often incredible

初步研究表明，一旦一个孩子向他们父母坦白，他们往往会不相信

Preliminary studies show that once a child confesses to their parents, they tend to not believe it

初步研究表明，一旦孩子向父母坦白, 他们通常都会被怀疑

Initial studies have shown that once children are confessed to their parents, they are usually skeptical

初步研究表明 一旦一个孩子像父母坦白，他们大多不会被信任

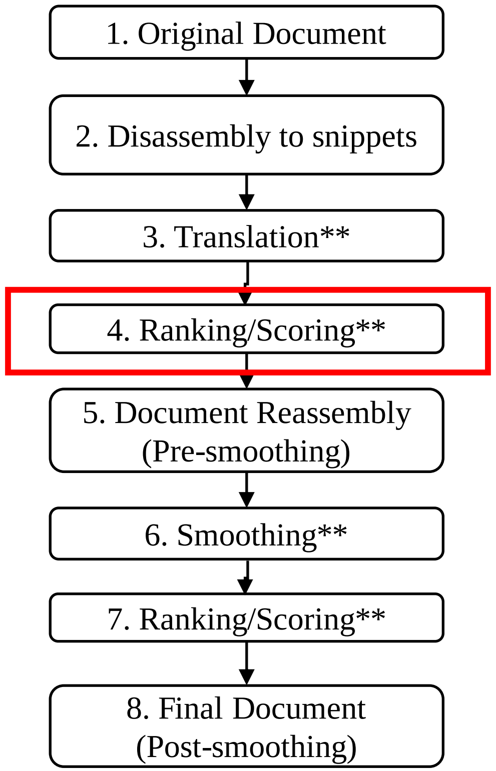
Preliminary studies show that once a child is confused as a parent, most of them will not be trusted

初步研究表示，一旦孩子向父母坦白，他們常常抱持懷疑

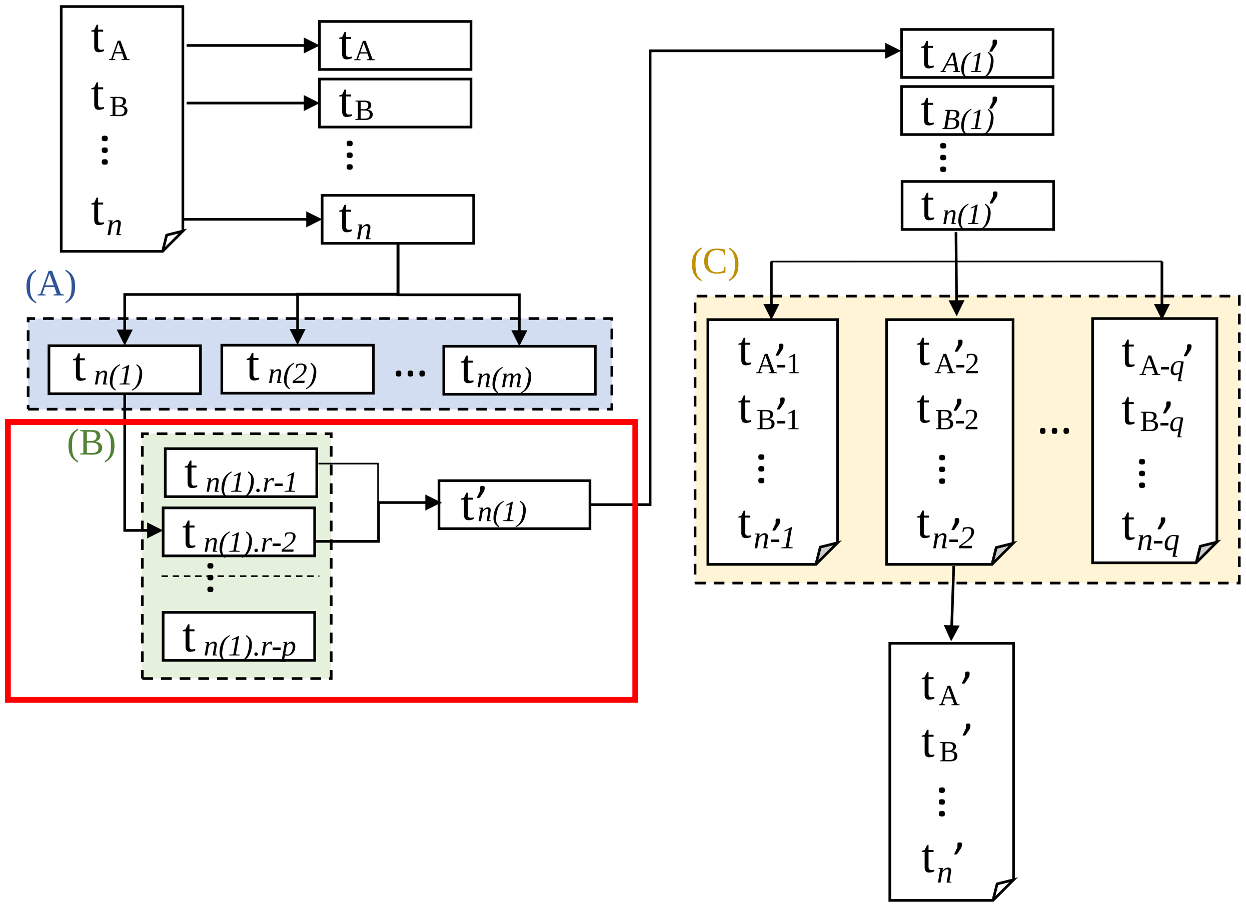
Initial studies show that once children are confessed to their parents, they often hold doubt

* = two identical translations were returned

Framework: Ranking or Scoring Translated Alternatives



1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)



** = crowd-assisted

Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

1 初步研究表明，一旦一个孩子向父母坦白，他们往往会被怀疑*
Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical*

4 初步研究显示，一旦孩子向父母坦白，他们往往难以置信
Preliminary studies have shown that once children are confessed to their parents, they are often incredible

2 初步研究表明，一旦一个孩子向他们父母坦白，他们往往会不相信
Preliminary studies show that once a child confesses to their parents, they tend to not believe it

5 初步研究表明，一旦孩子向父母坦白, 他们通常都会被怀疑
Initial studies have shown that once children are confessed to their parents, they are usually skeptical

6 初步研究表明 一旦一个孩子像父母坦白，他们大多不会被信任
Preliminary studies show that once a child is confused as a parent, most of them will not be trusted

3 初步研究表示，一旦孩子向父母坦白，他們常常抱持懷疑
Initial studies show that once children are confessed to their parents, they often hold doubt

* = two identical translations were returned

Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

1 初步研究表明，一旦一个孩子向父母坦白，他们往往会被怀疑*
Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical*

4 初步研究显示，一旦孩子向父母坦白，他们往往难以置信
Preliminary studies have shown that once children are confessed to their parents, they are often incredible

2 初步研究表明，一旦一个孩子向他们父母坦白，他们往往会不相信
Preliminary studies show that once a child confesses to their parents, they tend to not believe it

5 初步研究表明，一旦孩子向父母坦白, 他们通常都会被怀疑
Initial studies have shown that once children are confessed to their parents, they are usually skeptical

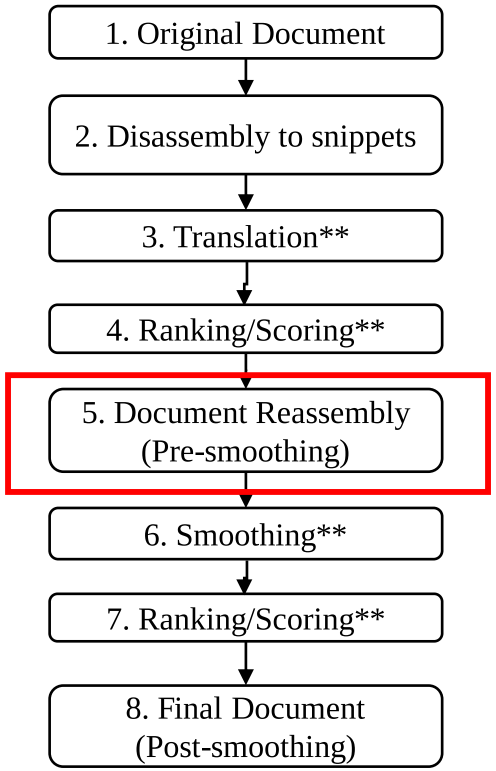
6 初步研究表明 一旦一个孩子像父母坦白，他们大多不会被信任
Preliminary studies show that once a child is confused as a parent, most of them will not be trusted

3 初步研究表示，一旦孩子向父母坦白，他們常常抱持懷疑
Initial studies show that once children are confessed to their parents, they often hold doubt

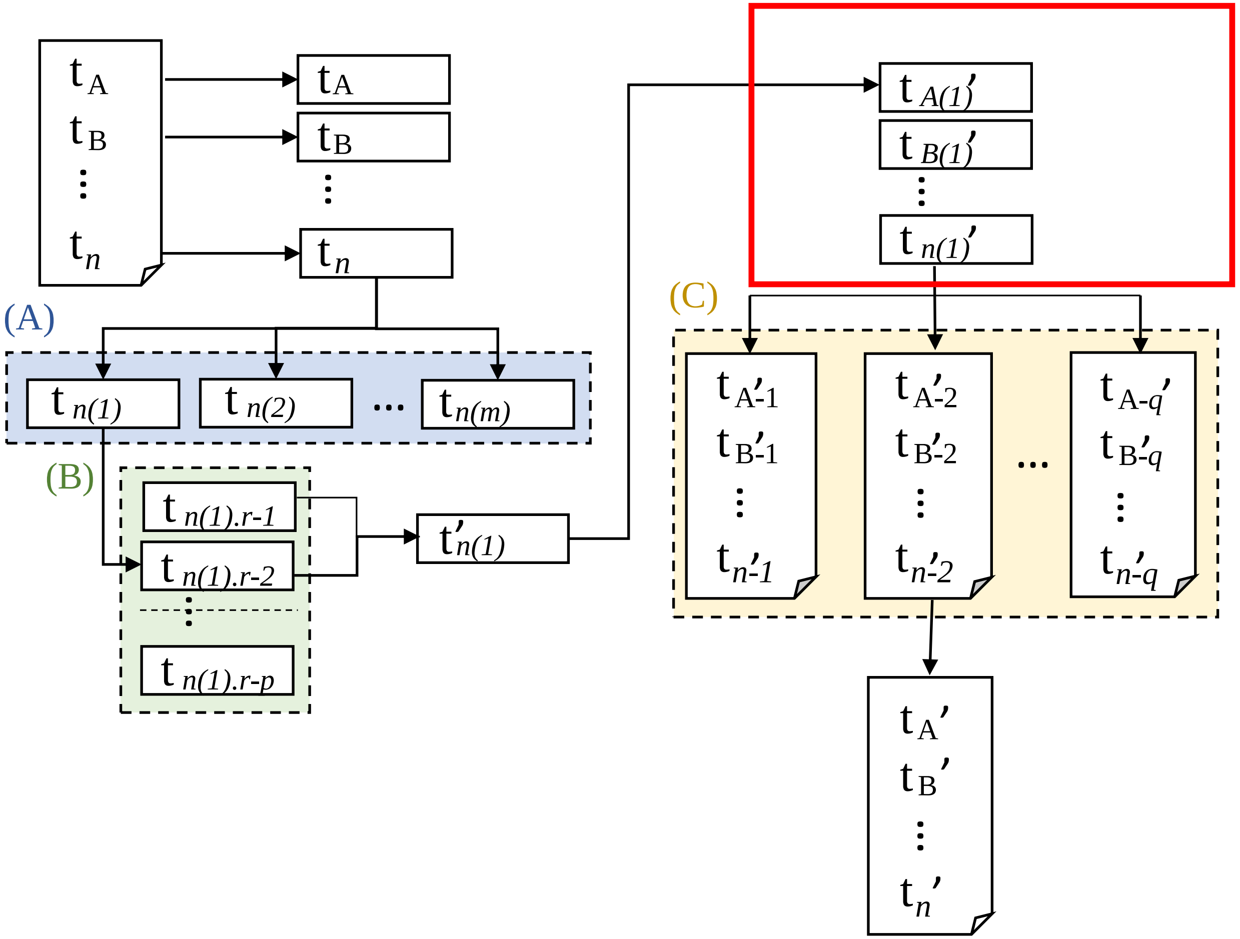
Normally we set a selection depth of 1 (winner take all), but if we set a selection depth > 1 (above we have n=3), we can actually have multiple translation-ranking cycles for each snippet

* = two identical translations were returned

Framework: Reassembly

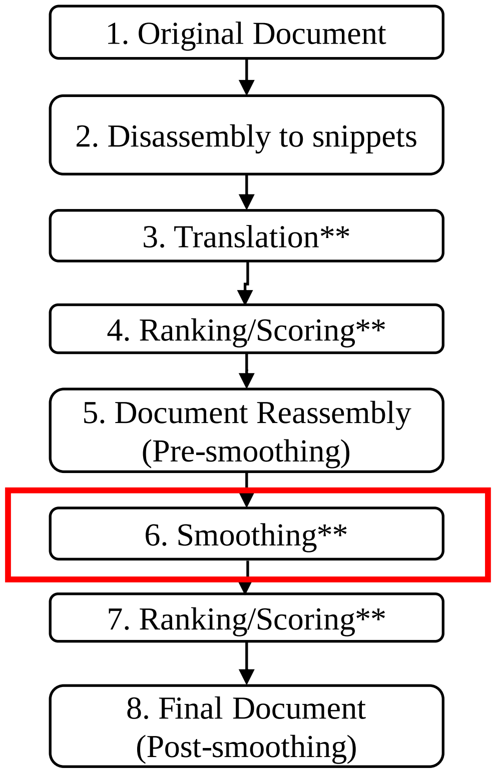


1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)

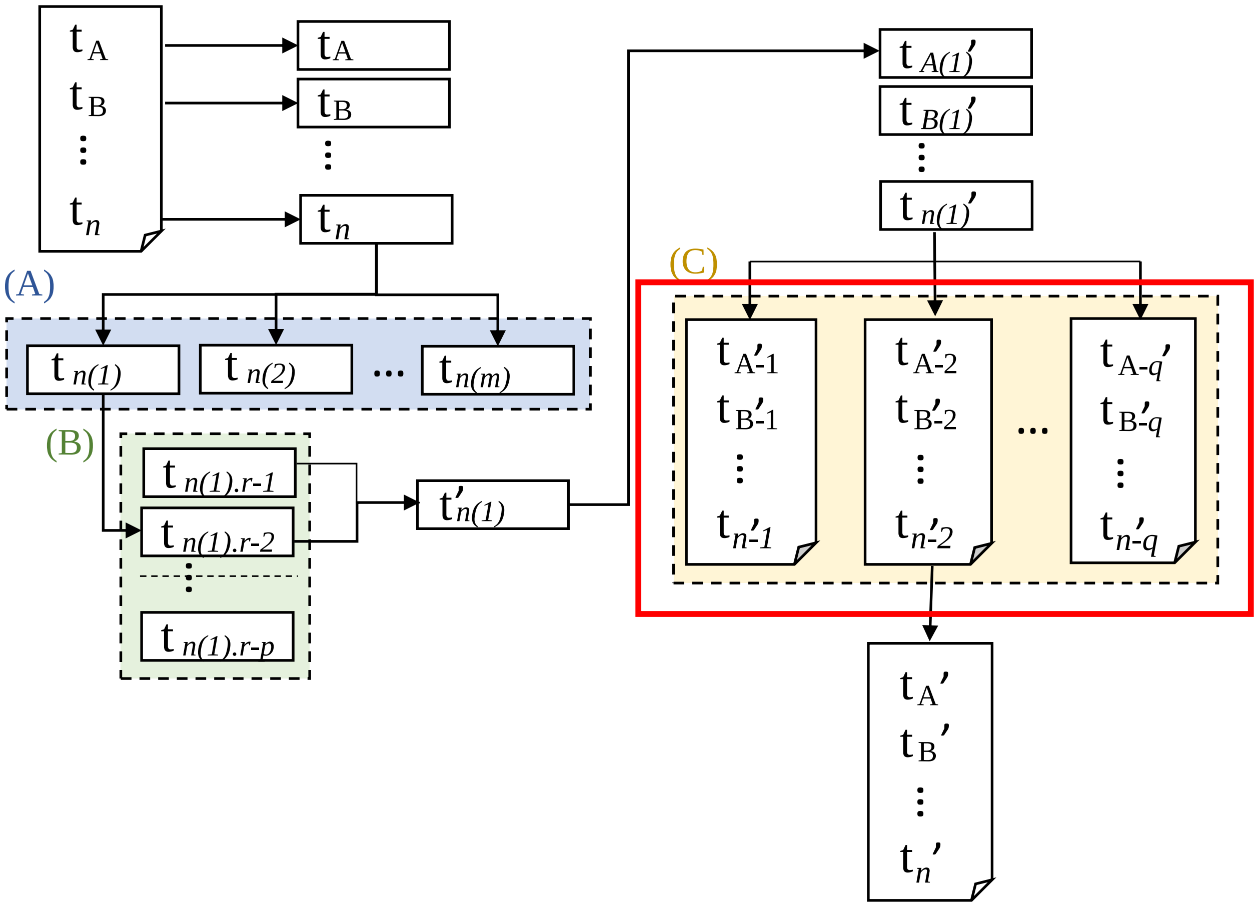


** = crowd-assisted

Framework: Smoothing/Editing



1. Original Document
2. Disassembly to snippets
3. Translation**
4. Ranking/Scoring**
5. Document Reassembly (Pre-smoothing)
6. Smoothing**
7. Ranking/Scoring**
8. Final Document (Post-smoothing)



** = crowd-assisted

Preliminary studies have shown that once a child confesses to their parents, they are often held in disbelief

初步研究表明，一旦一个孩子向父母坦白，他们往往会被怀疑

Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical

初步研究表明，一旦一个孩子向他们父母坦白，他们往往会不相信

Preliminary studies show that once a child confesses to their parents, they tend to not believe it

初步研究表示，一旦孩子向父母坦白，他們常常抱持懷疑

Initial studies show that once children are confessed to their parents, they often hold doubt

初步研究表明，一旦一个孩子向父母坦白，他们往往会被怀疑

Preliminary studies have shown that once a child confesses to their parents, they tend to be skeptical

Monolinguals Can Also Be Used as Text Editors/Smoothers

每个翻译软件都有不足，所以有时需要同时使用它们

From Google:
Each translation software is inadequate, so sometimes you need to use them at the same time

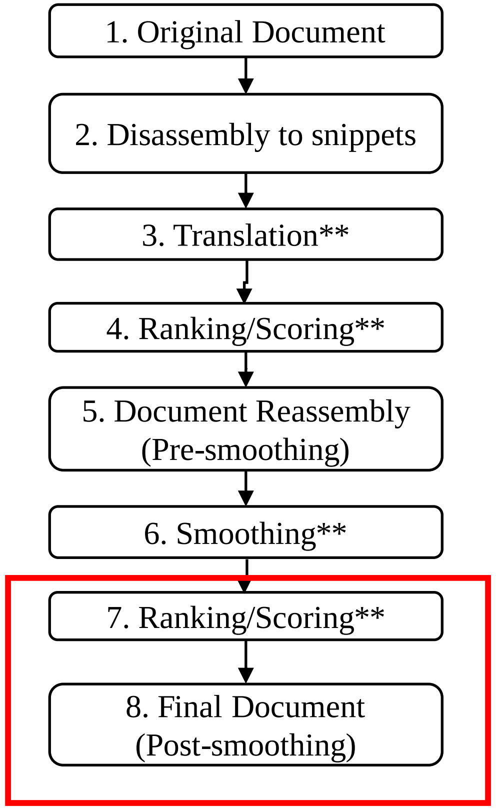
From Baidu:
Each translation software is not enough, so sometimes need to use them at the same time

From Youdao:
Every translation software has a deficiency, so it is sometimes necessary to use them simultaneously

From Bing:
Each translation software is deficient, so it is sometimes necessary to use them simultaneously

Each translation software tool has its own imperfections, so it is sometimes necessary to use more than one simultaneously

Framework: Ranking or Scoring the Edited Alternatives



1. Original Document

2. Disassembly to snippets

3. Translation**

4. Ranking/Scoring**

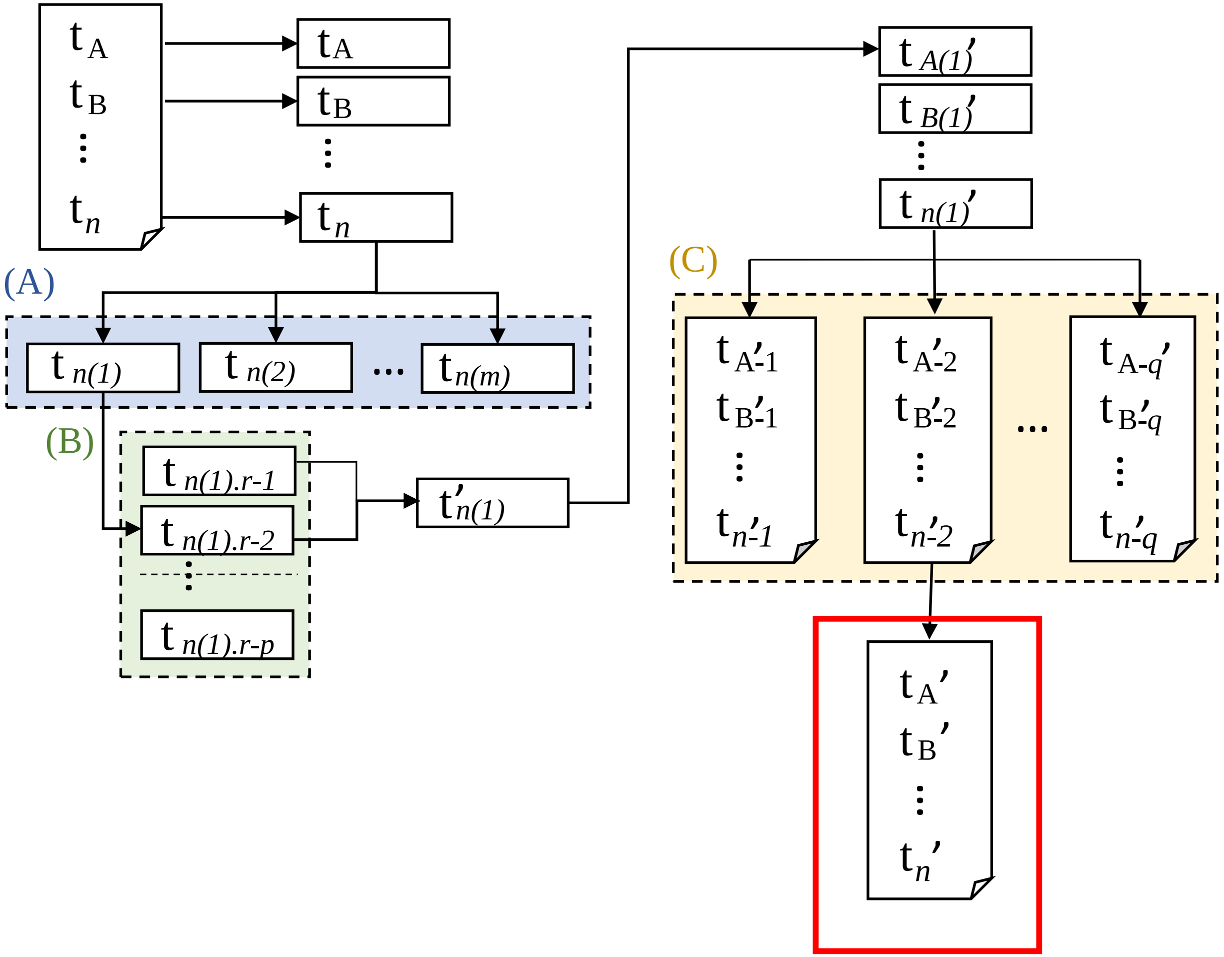
5. Document Reassembly (Pre-smoothing)

6. Smoothing**

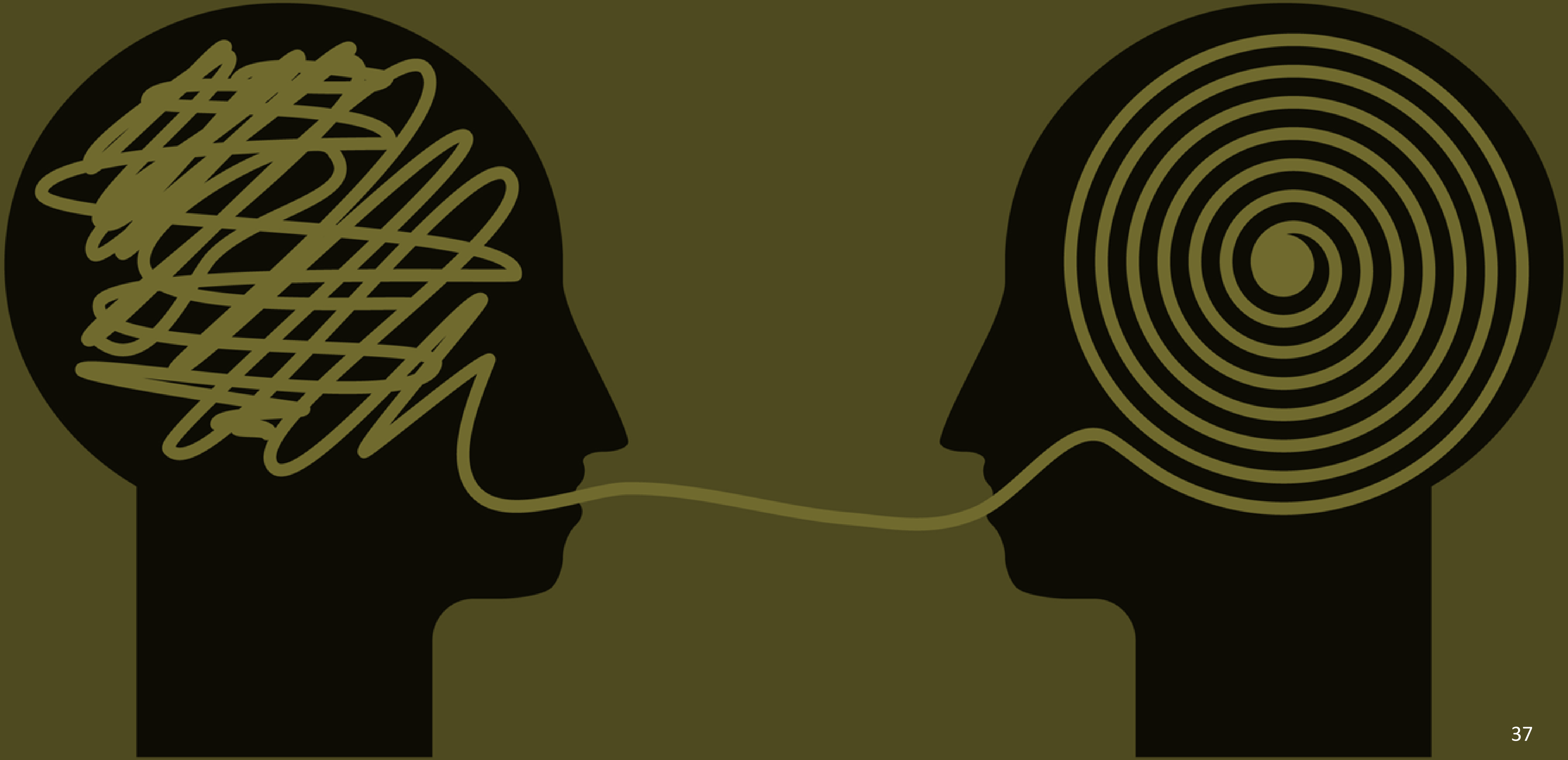
7. Ranking/Scoring**

8. Final Document (Post-smoothing)

** = crowd-assisted



Empirical Experiments



Pilot Study: Use Ranking (Voting) or Scoring (Ranking)?

- End result was similar if 7 items were to be ranked/scored
  - More than 7 → scoring
  - Fewer than 7 → ranking

- Stronger preference for ranking when:
  - Size of each snippet contained more text
  - Larger disparity between snippets

Experiment: Collections Used

English-to-Chinese translation

- Used the first 4 paragraphs of 10 randomly-selected OHSUMED articles (Hersh, 1994)

English text summarization

- Used the same 10 randomly-selected OHSUMED articles (Hersh, 1994)
- Limit of 10 words per snippet/40 words for the final smoothed summarization.

Experiment: Baselines

English-to-Chinese translation

- Professional: Paid a professional translator $186.30 for the 10 articles
- Crowd: 10 crowdworkers each translate one full document ($1.25 per document, $0.01/word) for a total cost of $12.50.

English text summarization

- Professional: $6.00 per summarization (an effective rate of $0.15/word), for a total of $60.00
- Crowd: $0.40 per summarization($0.01/word), for a total of $4.00.

Checking Quality: Language Tests

- Honeypots (obvious questions someone paying attention would notice)
- Check against known MT tools and eliminate those that are identical
  - But what if the MT tool happens to be identical to the translator?
    - Short snippets
    - Those with few variations

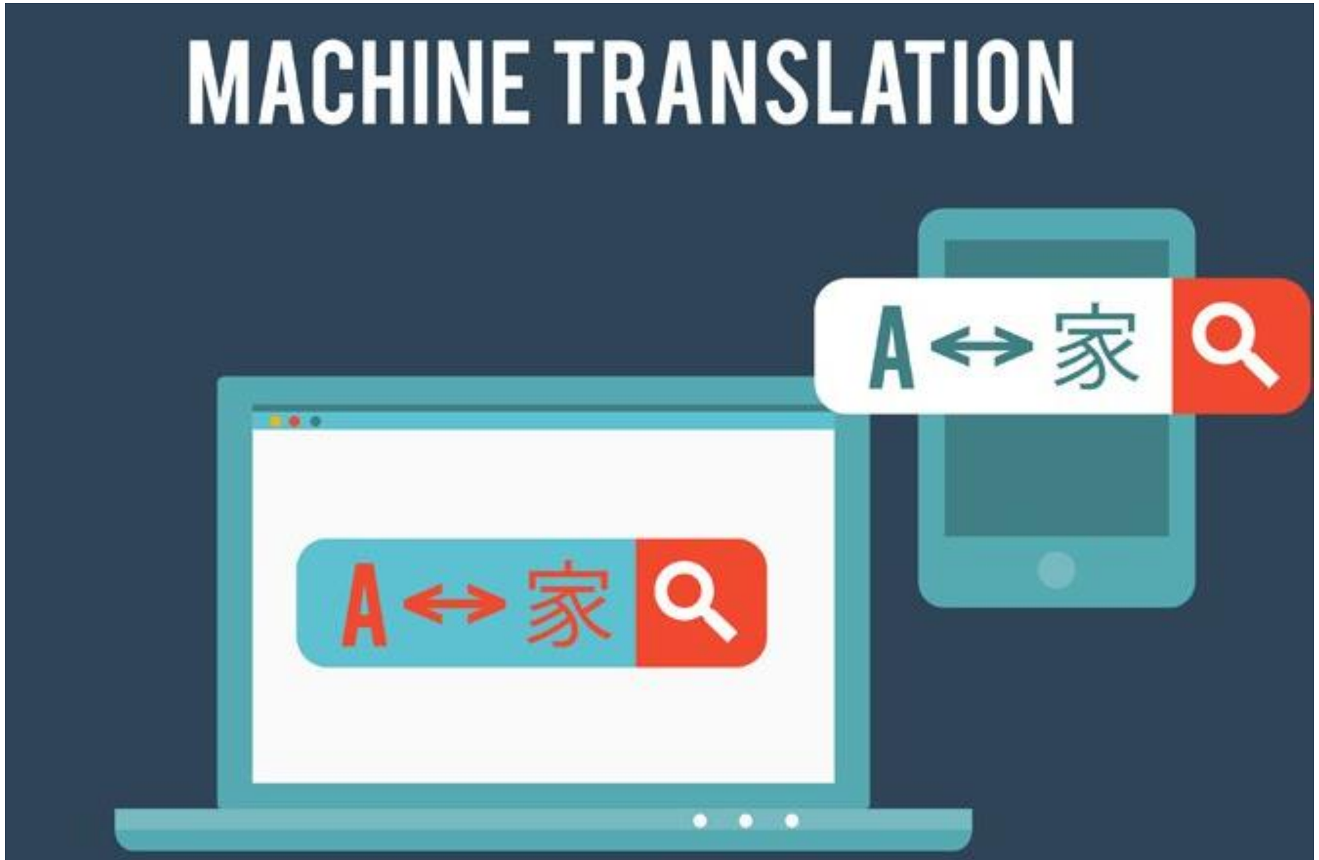
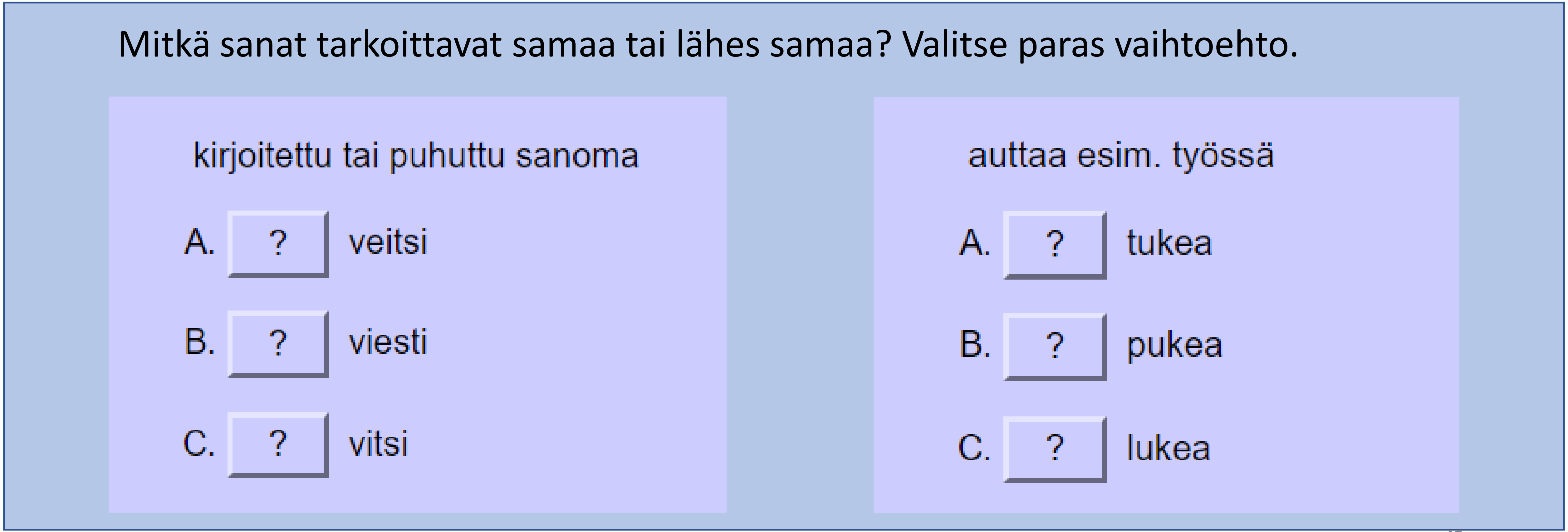


Image source: https://elearningindustry.com/4-machine-translation-tools-incorporating-machine-translation

Checking Quality: Language Tests

Known language tests, such as this one in Finnish



Mitkä sanat tarkoittavat samaa tai lähes samaa? Valitse paras vaihtoehto.

kirjoitettu tai puhuttu sanoma

A. ? veitsi

B. ? viesti

C. ? vitsi

auttaa esim. työssä

A. ? tukea

B. ? pukea

C. ? lukea

http://www.suomikoulut.fi/yki/sanastoharjoituksia7.htm

Experiment: Payments to the Crowd

English-to-Chinese translation

- Snippet translators
  - 7 translators (@$0.10 per snippet)
  - 3 rankers (@$0.05)
- Smoothers/editors
  - 7 translators (@$0.10 per document)
  - 3 rankers (@$0.05)

English text summarization

- Snippet text summarizers
  - 7 summarizers (@$0.10 per snippet)
  - 3 rankers (@$0.05)
- Smoothers/editors
  - 7 summarizers (@$0.10 per document)
  - 3 rankers (@$0.05)

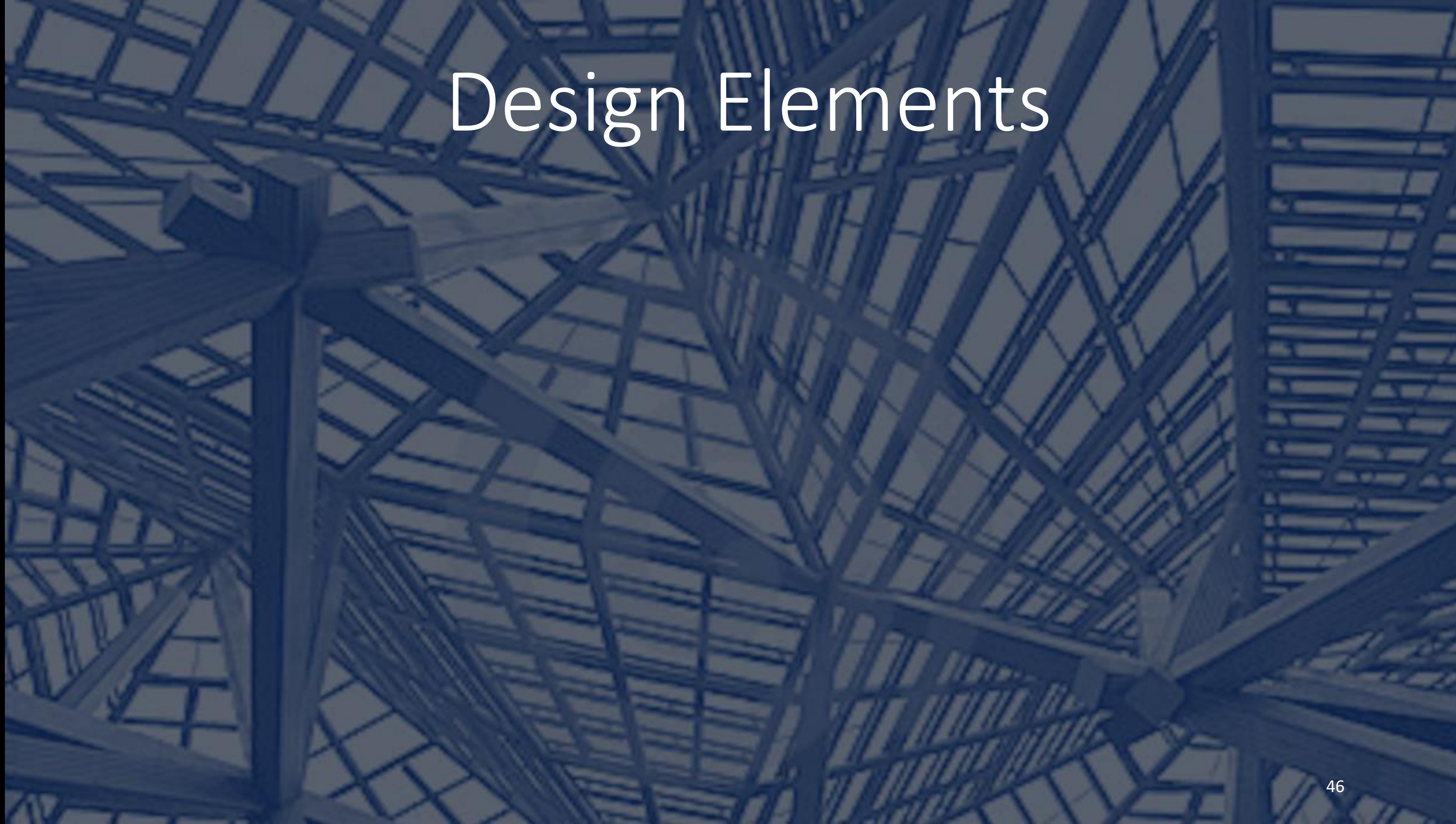
Experiment: Translation Results

Translation	Pre-smoothing		Post-smoothing	
EN to CH	BLEU	Time	BLEU	Time
Google Translate API	--	--	32.38	0:01
Baseline - Professional	--	--	40.54	29:01
Baseline - CS	--	--	29.18	6:08
CS First 1	21.44	6:20	28.9	5:53
CS First 3	23.02	9:48	35.71	8:26
CS First 5	27.93	10:12	38.65	10:29
CS All 7	29.74	13:23	39.81	12:36

Experiment: Text Summarization Results

Summarization	Pre-smoothing		Post-smoothing	
	BLEU	Time	BLEU	Time
TextRank*	--	--	34.46	0:01
Baseline - Professional	--	--	44.61	6:16
Baseline - CS	--	--	38.98	2:08
CS First 1	32.33	1:03	36.42	1:19
CS First 3	36.02	2:59	42.29	2:13
CS First 5	37.15	3:54	43.61	3:30
CS All 7	38.96	4:49	45.95	5:14

* = http://summanlp.github.io/textrank



Design Elements

Evaluation of Cost – Translation Task

Number of CS workers Used for each step of translation task	Difference in BLEU score between professional translator and score achieved with this number of CS workers	Total cost of using this number of CS workers for translation	Difference in cost between using a professional translator ($186.30) and the CS workers	Amount paid for each 1 additional BLEU point using a professional over this number of CS workers
1	11.64	$ 7.00	$ 179.30	$ 15.40
3	4.83	$ 11.00	$ 175.30	$ 36.29
5	1.89	$ 15.00	$ 171.30	$ 90.63
7	0.73	$ 19.00	$ 167.30	$ 229.18

Evaluation of Cost – Text Summarization Task

Number of CS workers Used for each step of summarization task	Difference in BLEU score between professional and score achieved with this number of CS workers	Total cost of using this number of CS workers for summarization	Difference in cost between using a professional ($60.00) and the CS workers	Amount paid for each 1 additional BLEU point using a professional over this number of CS workers
1	8.19	$ 7.00	$ 53.00	$ 6.47
3	2.32	$ 11.00	$ 49.00	$ 21.12
5	1.00	$ 15.00	$ 45.00	$ 45.00
7	-1.34	$ 19.00	$ 41.00	$ (30.60)

Evaluation of Time – Translation Task

Number of CS workers Used for each step of translation task	Difference in BLEU score between professional translator and score achieved with this number of CS workers	Number of hours taken for translation with this number of CS workers	Difference in time taken, in hours between the professional and CS workers	Number of additional hours needed to increase the BLEU score 1 point with this number of CS workers
1	11.64	12.22	16.80	0.69
3	4.83	18.23	10.78	0.45
5	1.89	20.68	8.33	0.23
7	0.73	25.98	3.03	0.24

Evaluation of Time – Text Summarization Task

Number of CS workers Used for each step of the summarization task	Difference in BLEU score between professional and score achieved with this number of CS workers	Number of hours taken for summarization with this number of CS workers	Difference in time taken, in hours between the professional and CS workers	Number of additional hours needed to increase the BLEU score 1 point with this number of CS workers
1	8.19	2.37	3.90	2.10
3	2.32	5.20	1.07	2.18
5	1.00	7.40	(1.13)	(0.88)
7	-1.34	10.05	(3.78)	0.35

Conclusions and Future Work

Evaluate qualities

✓ 1

Robust:

- Our framework should be impervious to low-quality inputs from a malicious crowdworker.

✓ 2

Verifiable:

- Should be able to perform an evaluation of outputs after each crowdworker-dependent step in our framework.

✓ 3

Consistent:

- The same inputs should produce approximately the same outputs, even with different crowdworkers.

✓ 4

Flexible:

- As few components as possible should rely exclusively on multi- and bilingual crowdworkers.

Conclusions

Developed a framework

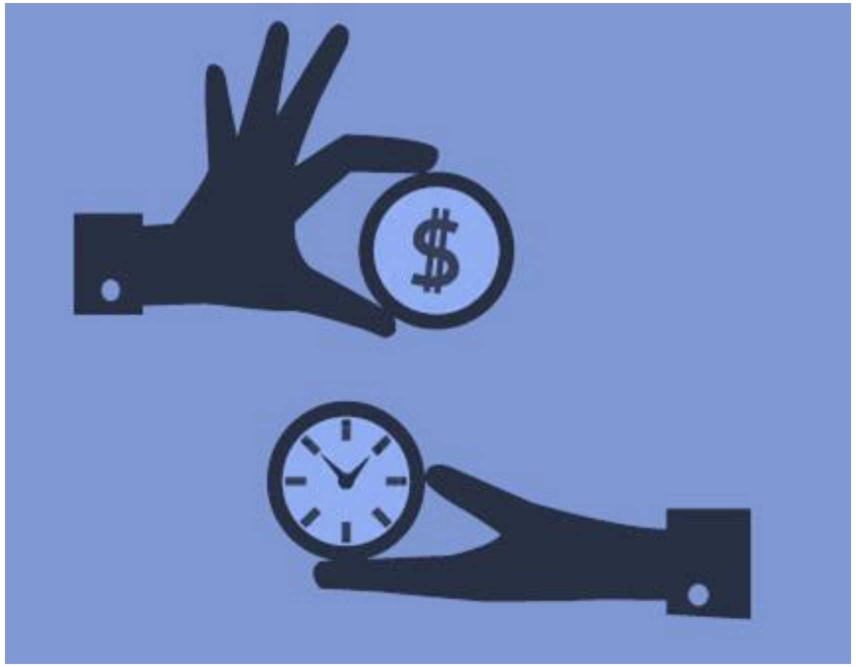
- Smoothing step really helps!
- Found 3-5 crowdworkers can produce very good results
- Beyond 5 crowdworkers really does not affect our results much

…but this is a small study… more needs to be done

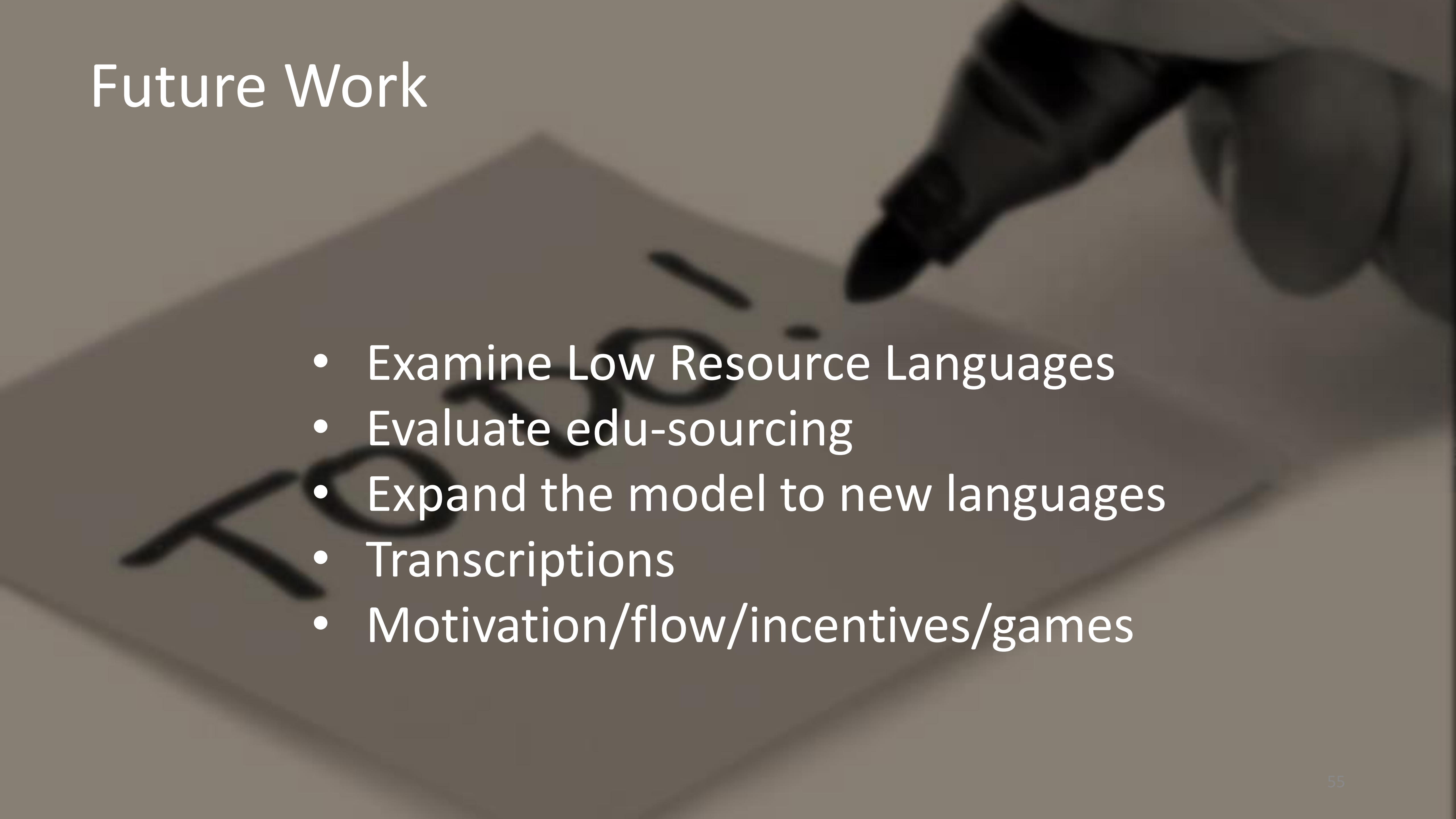
Conclusions

From initial appearances, it is very cost- and time-effective

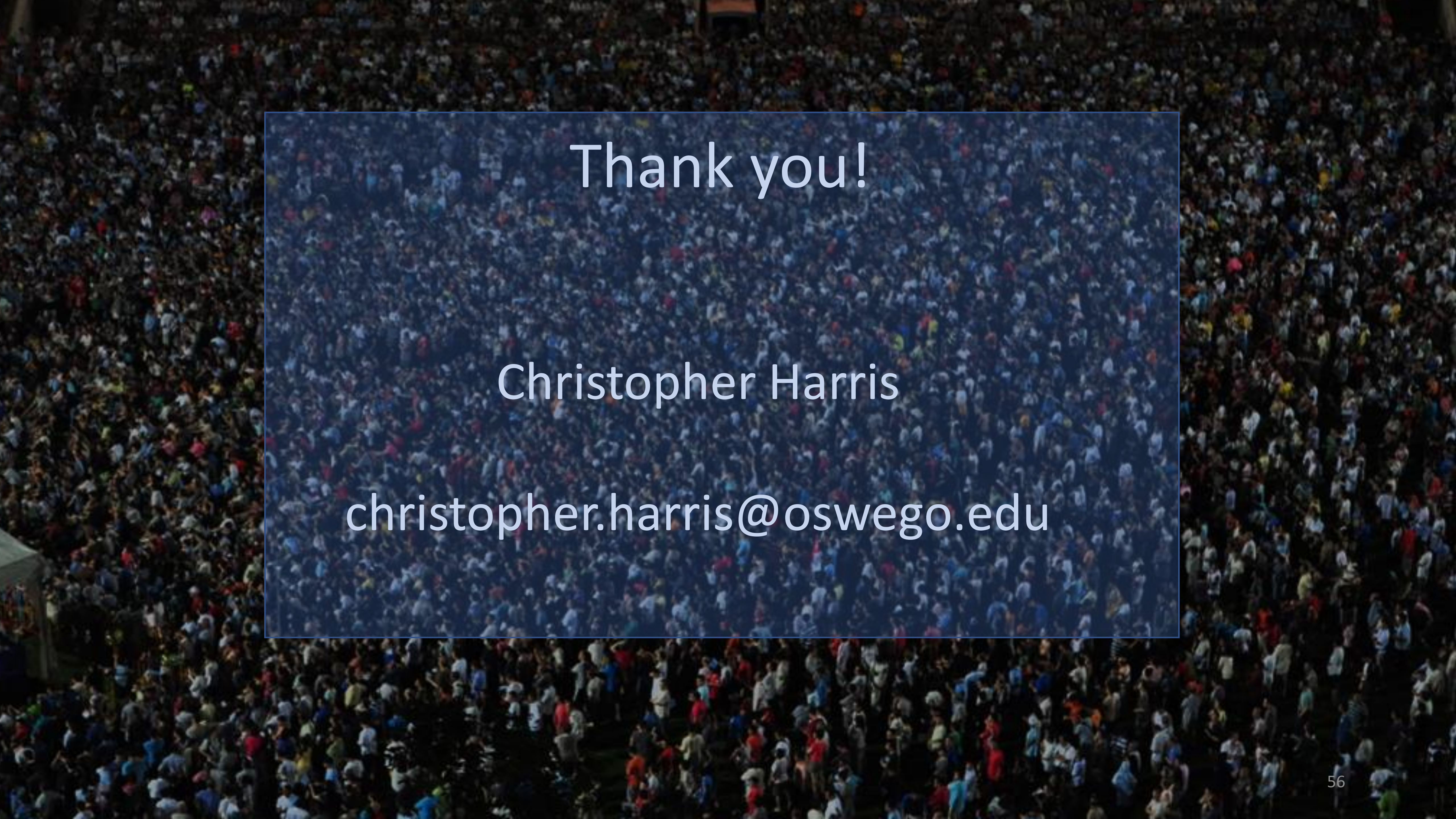
Task	Metric	Relative to Professionals	What others have experienced
Translations	Cost	1/20th (using 5 + 3 workers)	1/23rd (Harris & Xu, 2011), 1/30th (Callison-Burch, 2009)
	Time	1/3rd	N/A
Text Summarizations	Cost	1/4th (using 5 + 3 workers)	N/A
	Time	1/6th	N/A



Future Work



- Examine Low Resource Languages
- Evaluate edu-sourcing
- Expand the model to new languages
- Transcriptions
- Motivation/flow/incentives/games



Thank you!

Christopher Harris

christopher.harris@oswego.edu