Multilingual and Parallel Corpora
Evaluation

Amir Zeldes

amir.zeldes@georgetown.edu

Chatbots and MT - discussion

- Are chatbots doing a kind of machine translation?
  - Why?
  - Why not?
  - If not, what does that mean about what MT is doing?
- What are the implications for Artificial Intelligence?
- Is Deep Learning the future?

http://neuralconvo.huggingface.co/

AI and parallel data: Some insidious challenges

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal sword in hand:
Long time the manxome foe he sought—
So rested he by the Tumtum tree,
And stood awhile in thought.

And as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

One, two! One, two! and through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went galumphing back.

"And hast thou slain the Jabberwock?
Come to my arms, my beamish boy!
O frabjous day! Callooh! Callay!"
He chortled in his joy.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

Jabberwocky / Lewis Carroll

- How do you translate this?

Some insidious challenges

- French: (literary)

- Il brilgue: les tôves lubricilleux
  Se gyrent en vrillant dans le guave.
  Enmîmés sont les gougebosqueux
  Et le mômerade horsgrave.

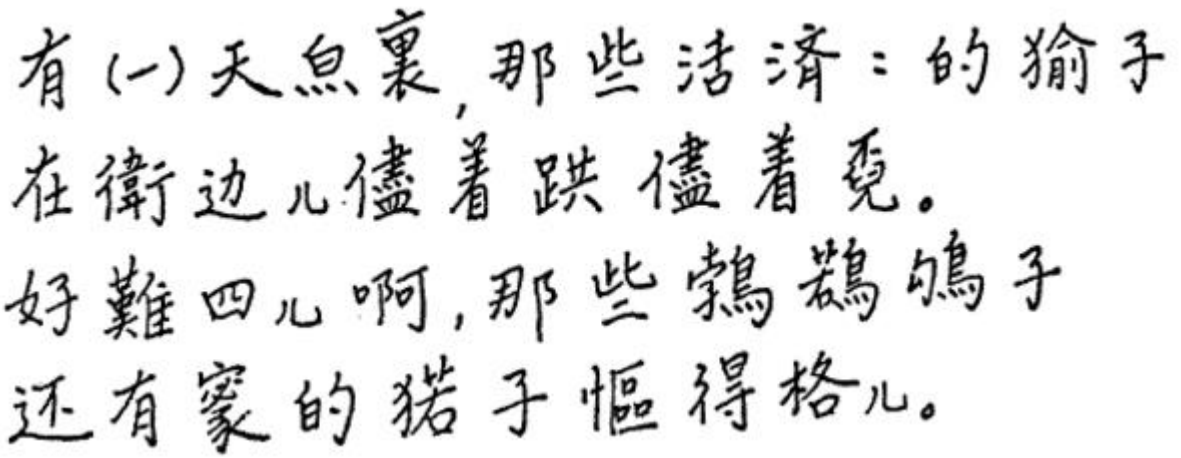
Some insidious challenges

- French: (Google Translate)

- Il était brillant, et les toits fendus,
  gyre et gimble dans le wabe,
  tous les mimsy étaient les borogoves,
  et le môme raths outgrabe.

Some insidious challenges

- Chinese



Yuen Ren Chao

And later on, when Humpty Dumpty explains the etymology of the difficult words, it will of course have to come out right in the translation. For example, "in the wabe" is translated as *tzay weybial*, since just as "wabe" comes from "way before," "way behind," and "way beyond," so does *weybial* come from *jeybial*, *neybial*, and *waybial*, that is, "this side," "that side," and "outside."

Some insidious challenges

- Hebrew
- הִבְרִיל כְּבָר,
  זַחְלָצִים קְלִיחִים
  חָגְווּ וְעָגוּ בַּשְׁבִּילֵל,
  מַסִּים הָיוּ הַסְמַרְלַחִים
  וְחֶזְרוֹנִי צִרְלֵל.

Next topics

- Today - Evaluation
- Language comparison
  - Special types of alignment
  - Parallel treebanking and complex phrase alignment
- Semantic mirror and multilingual lexicography

➢Please think about final paper topics!

How can we tell if a translation is good?

- Bad translations can have consequences:
  - http://news.nationalpost.com/news/world/bad-translation-of-zidong-jiashi-to-blame-for-chinas-first-tesla-autopilot-crash-driver-says
- How can we quantify? Recall Greenwich 1 vs 2/3:
  - Classify shifts
  - Count:
    - Total shifts
    - Some types more acceptable than others
    - Sensitivity to omission
- Can we automate evaluation metrics on parallel data?

Reference translation

- MT system performance is evaluated on test sets
- We reserve a limited amount of the parallel data
- Use human translation(s) as a reference:
  - Did the system produce the reference translation?
  - Or something close?

➢ Assuming output is not identical to the reference, how can we measure quality?

Thought exercise

- How would you make an evaluation metric that is:
  - Objectively decidable
  - Automatically calculable
  - Correctly penalizes broken translations
  - Correctly prioritizes translations by quality
  - Accepts alternative translations

- And ideally: try to come up with a number between 0-1

Thought exercise

- Source sentence: ***Il y a un chat sur le tapis***

- Reference translations:
  - There's a cat on the mat
  - A cat is on the mat
- Candidates:
  - The cat is on the mat
  - There's a dog on the mat
  - The cat
  - The the the the the the

MT evaluation metrics

- The most widespread metrics:
  - BLEU (Papineni et al. 2002)
  - ROUGE (Lin 2004)
  - METEOR (Banerjee & Lavie 2005)

- Attempt to solve precisely this exercise!

BLEU

- **BLEU (bilingual evaluation understudy)**
  - Bounded between 0 and 1
  - Supports **multiple** reference translations
  - Based on **precision:** how many of the TT words are actually in the reference translation?
- But raw precision is a bad metric:

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

- Perfect!

BLEU

- Fixing the problem:
  - We cap each word's allowed occurrences at max(m) appearances in any reference translation (the -> 2)
  - Take proportion of summed *m*'s out of candidate length:
    - the the the the the the the -> 2/7
- Problems:
  - Bias favors short translations:
    - the the -> 1/2
    - the cat -> 2/2!
  - No word sequence/ordering -> correctable by adding metrics for n-grams (in practice, up to n=4)

BLEU

- Additional refinements:
  - Also factor in **recall** (mat, is, on -> not recalled)
  - Can be problematic in the face of **many** reference translations
  - System should produce a sentence containing all words??
    - The cat is on the mat
    - There is a kitty on the mat
    - **Good candidate:** there the cat kitty is on the mat
  - Add a **brevity penalty** based on proportion of reference corpus size to output corpus size

Homework

- Translate by Friday:
  - On the subject at hand , I think that the people of Europe must be able to be confident that the goods - however dangerous they are - which are transported on Europe 's roads , railways , and so on are as safe as possible .

Homework

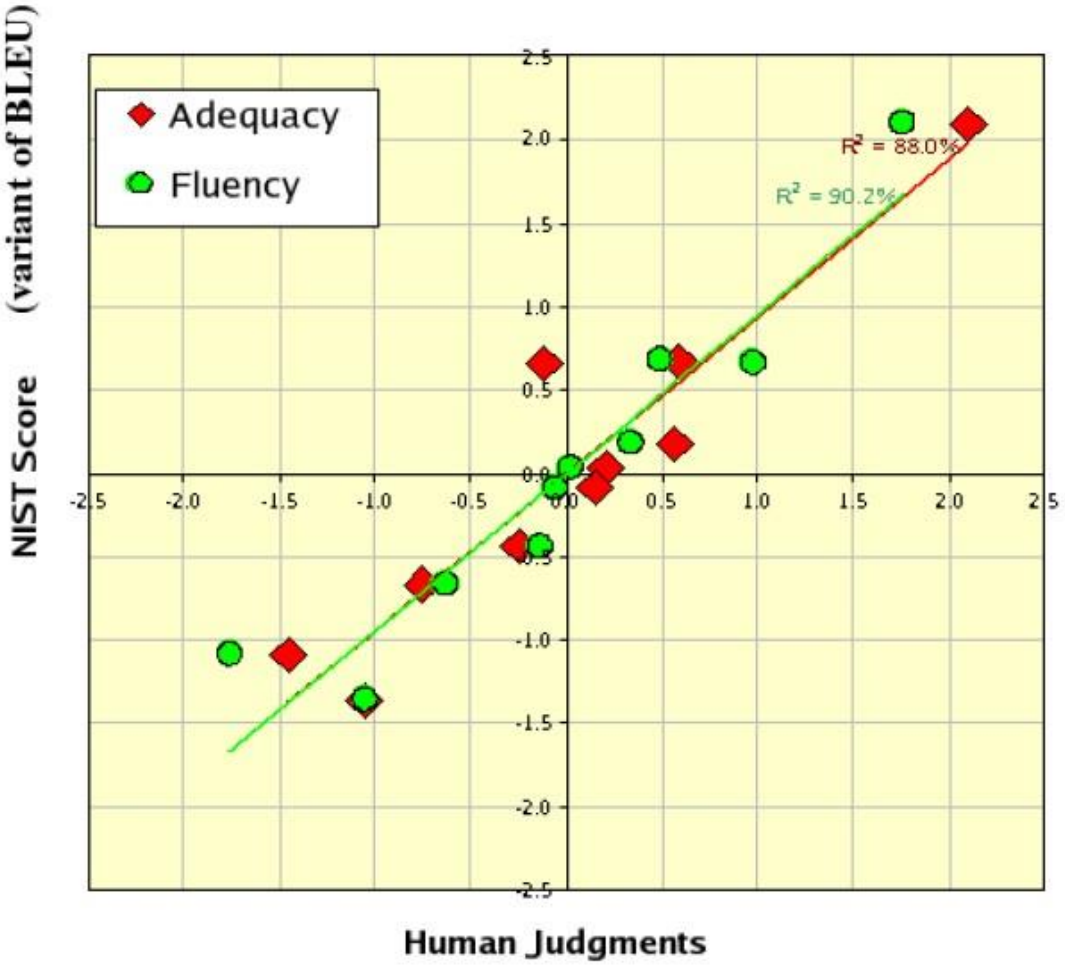
- Calculate BLEU-4 (1,2,3,4 gram precision), report length diff
- Deu: xxx xxxxx x xxx xxxxx x xxx xxxxxx xxxxxxx xxxxxx xxxx xxxxxx xxxxxxxxx xxxxxx x xxxx xxx x xxx xxx xxxxxxx xxxxxxx x xxxxxxxx xxxx xxxxxxxxxxxxx xxxx x xxxx xx xxxx xxxx xxxx xxxxxxxxxxx xxxxx xxxx x xx xxxxxx xxx xxxxxxx xxx x
- Spa: xxxxx xx xx xxxx x xxxx xxx xxx xxxxxxxxxx xx xxxxxx xxxxxx xxxxxxx xx xxx xx xxx xx xxxxxxxxxx xx xxxxxx xxx xxxxxxxxx x xxx xxxxxxxxxx x xxx xxx xxxx xxxxxxxxxx xx xxxxxxxxxx xxx xxxx xx xxxxxxxxx xxxxxxx x xxx xxxxxx xxxxxxxxx
- Fra: xxxx xx xxx xxx xx xxxxx xx xxxxxx xxxxxxx xxxxxxx xxxxxxx xxx xx xxxx xxx xxx xxxxxxxxxx xxxxxx xx xxxxxx xxxxx xx xxxxxxxx xxxxxxxxxx x xxx xx xx x xxxx xxxxxxxxxxx xx xxx xxxxxxx xxxxx xxxx xxx xxxxxxx
- Ita: xxxxxxxx xxx xxxxxx x xxxxx xxxxx xxx x xxxxxxxxx xx xxxxx xxxxx xxxxx x xxxxxxxxx xx xxxxxxxxxx xx xxxxxxx xxxxxxxxx x xxxxx xxx xxxx xx xxx xx xxxxxx xx xxxxx xxxxxxxxxx x
- Rus: xx xxxxxxxxx x xxxxxx xxx xxxx xxxxxx xxxxxx xxxx x xxxxxxxxx xxxxxxxxxx x xxxx xxx xxxxxxx xxx xx xxx xxxxxxx xxxxxxxxxxxxxxxx xx xxxxxxx xxxxxxx xxxxxxxx xxxxxxx x xx xxx xxxxxxxxx xxxxxxxxxx xxxxxxxxx xxx xxxxxxxx
- Zho: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx x xxxxxx x xxxxxxx
- Kor: xx xxxx xxx xxxx xxx xxxx xxx xxx xx xxx xxxx xxx xxx x xxxxx xx x x xxxxxx xxxxxx

Reveal Friday at noon!

BLEU

- Is this a good approximation of human judgment?

Correlation with human judgment?



- (Doddington 2002, repr. In Koehn 2011)

BLEU

- Is this a good approximation of human judgment?
- Problems at the individual sentence level:
  - No accounting for **synonymy**:
    - Reference: Fred hit the deer with his truck
    - C1: Fred ran over the deer with his truck
    - C2: Fred vaporized the deer with his truck ←
  - Vulnerable to morphosyntactic paraphrase:
    - I'm hungry
    - Ich habe hunger <> ich bin hungrig

Just get more references translations?

- NIST 2001 data set (from Koehn 2011)

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

- But: Number of **reference translations** alters score

BLEU

- But at the corpus level, problems should(?) even out
- De facto standard for about a decade

- What happens when systems optimize towards BLEU rather than ‘subjective good translation’?