(Machine) Translation Evaluation

Maja Popović

ADAPT Centre,
School of Computing
Dublin City University

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

Engaging Content
Engaging People

Overview

Notion of MT quality

Why is translation evaluation not trivial?

- there are many ways to say the same thing in a language
- there are many ways to translate the same thing into another language

Example: any good translations here?

source text:	Es wird eine Art von Brücke sein.
human (reference) translation:	It will be a sort of bridge.
MT system 1:	It is almost as a bridge act.
MT system 2:	It will act as a bridge.
MT system 3:	It will not act as a bridge.
MT system 4:	It will sort of bridge be.

- Is any of these translations a “good translation”?
- Why?

Possible goals of MT

- publishing
  (using directly without any revisions and corrections)
- post-editing
  (correcting errors)
- using the translation for another application
  (e.g. text classification, information retrieval, etc.)
- creating training data
  (for MT “back-translation”, or other NLP task)

Publishing

- system 2 (*It will act as a bridge.*) is the only usable option
  - preserves the meaning
  - easily readable/understandable
  - grammatically correct
- system 4: *It will sort of bridge be.*
  preserves the meaning, but grammar is bad
- system 1: *It is almost as a bridge act.*
  meaning kind of there, bad grammar, difficult to read
- system 3: *It will not act as a bridge.*
  perfectly readable and grammatical, but meaning completely changed

Post-editing

- system 2 is again the best
  nothing has to be corrected

- system 3 is the next best
  although it contains a critical error (negation particle “not”)
  deleting one word is an easy correction

Using translation for another application

- system 2 is again definitely the best
- system 4 is also acceptable
  meaning is preserved, grammar is not critical
- even system 1 could be acceptable
  because the meaning is still there
- system 3 is not acceptable
  because the meaning changed by negation

Another example to think about

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.

What does “translation quality” mean?

- Meaning preserved? (adequate)
- Good grammar? (fluent)
- Readable?
- Easy to correct (post-edit)?
- Usable? What for?
- Some of those?
- All of those?

Quality for whom/what?

- end users (publishing, gisting)
- post-editors
- other applications (text classification, IR)
- MT system developers (system abilities and possibilities for improvement)

Manual evaluation

Main quality aspects

Main quality aspects

- the definition of quality depends on the task/goal
- still, in a manual evaluation, some of the following three aspects will always be taken into account:
  - adequacy
  - fluency
  - readability

Adequacy

Adequacy (accuracy, meaning preservation, fidelity)

describes how well the **meaning** of the original text is conveyed in the translated text

- goal of translation:
  to bring the meaning of a text in one language into another language

⇒ **if the meaning is not preserved, the translation did not fulfill its purpose**

the evaluators must see the original text in the source language
(or a reference translation in the target language *(not recommended)*

Fluency

Fluency (grammaticality)

describes the **grammar** of the target language in the translated text

- not crucial
- however, readers (users, evaluators, translators, ...) like fluent texts very much
- so much that adequacy errors are often overlooked/“forgiven” in fluent texts

the original source language text is not necessary, but it can be seen

usually assessed together with adequacy

Comprehension

Comprehension (readability, understandability)

describes the extent to which the reader is able to understand the translated text

- goal of translation:
  to enable understanding something written in another (unknown) language

⇒ **if a reader cannot understand the translated text, the translation did not fulfill its purpose**

the evaluators **must not** have access to the original text in the source language (nor to a reference translation in the target language)

Comprehension $\neq$ Fluency

Two perfectly fluent sentences:

Colourless green ideas sleep furiously.

Don't understand them if you hit a heavy suitcase.

Can you understand what they mean?

Usage of different criteria

- adequacy
  the most important and the most evaluated
- fluency
  the least important and frequently evaluated
- comprehension
  important but rarely evaluated

What is wrong with comprehension?

although important, it is rarely evaluated

probable reason:

1. adequacy and fluency are usually evaluated simultaneously,
   with the source text presented
2. adequacy and comprehension cannot be (properly) evaluated simultaneously
   - comprehension
     the source text **must not** be presented
   - adequacy
     the source text **must** be presented
3. evaluating only comprehension without adequacy might be dangerous

The danger of fluent readable *inadequate* translations

- readers prefer fluent and readable translations (totally normal)

!!! often, adequacy errors are overlooked/“forgiven” in those translations

Methods for manual evaluation

How is the evaluation performed?

- scoring (the mostly used method)
  assign an overall numerical score to each sentence/segment of a translation
- ranking
  compare two or more translations by ordering each sentence/segment from 'best' to 'worst'
- error marking
  find and mark (highlight) all problematic parts of a translation
- error classification
  find and classify (mark and assign an error label) all problematic parts of a translation

Scoring

assigning a number to the evaluated translation

- higher number ⇔ better translation
  - discrete values (usually from 1 to 5)
  - continuous values (usually from 0 to 100)

the assigned number refers to what?

to the selected criterion
usually: adequacy+fluency

Scoring from 1 to 5: example

possible scores for our example:

translation	text	Adequacy	Fluency	A+F	Compr.
reference	It will be a sort of bridge.	5	5	5	5
system 1	It is almost as a bridge act.	2	1	3	3
system 2	It will act as a bridge.	5	5	5	5
system 3	It will not act as a bridge.	1	5	2	5
system 4	It will sort of bridge be.	5	2	4	4

Ranking

ordering (ranking) translations from best to worst

- comparing pairs of translations is the easiest,
  but requires more comparisons in total
- often: comparing 5 translations
- ties are allowed

ordering translations according to what?

usually: a “general” unspecified criterion:
a mixture of adequacy and fluency, and comprehensibility to some extent

Ranking: example

possible order of translations:

translation	text	rank
reference	It will be a sort of bridge.	1
system 2	It will act as a bridge.	1
system 4	It will sort of bridge be.	2
system 1	It is almost as a bridge act.	3
system 3	It will not act as a bridge.	4

Scoring vs. Ranking

- a point for ranking:
  easier and faster (especially if only two translations are compared)

- a point for scoring:
  the information about the **actual quality** of each translation is lost in the ranking process

  What does ”rank 1” exactly mean?
  - ”a really good translation”?
  - or ”the least bad translation”?
  - or... (many possibilities)

Error marking

mark all errors in the translation

- lower error count ⇔ higher quality

what is considered as error?

error according to some criterion (e.g. adequacy)

absence of explicit criterion implies adequacy + fluency
(+ a little bit of comprehension)

Error marking: example

Highlighted errors in our example might look like this:

system	text
reference	It will be a sort of bridge.
system 1	It is almost **as** a bridge **act** .
system 2	It will act as a bridge.
system 3	It will **not** act as a bridge.
system 4	It will **[]** sort of bridge **be** .

Error marking vs scoring and ranking

- points for marking:
  - the information about the **actual errors** is lost in the scoring process
    (and even more in the ranking process)
- points for scoring/ranking:
  - ranking is faster
- still missing:
  - the information about the type of each error
    (Is it incorrect lexical choice? Or addition? Maybe case? Gender? Order? ...)

Error classification

mark and classify each error in translation

lower error count ⇔ higher quality

what is considered as an error?

- error types are described in a pre-defined error scheme

Error classification: example

reference:	It will be a sort of bridge.
system 1:	It **is**/*tense* **almost**/*addition* as a bridge **act**/*order*.
system 2:	It will act as a bridge.
system 3:	It will **not**/*addition* act as a bridge.
system 4:	It will **[]**/*omission* sort of bridge **be**/*order*.

Error classification vs error marking

- a point for error classification:
  - provides details about actual errors and problems
  - ± allows less freedom and subjectivity

- a point for error marking:
  - classification requires increased effort and time
  - pre-defining a scheme is not trivial
  - ± potential bias towards the scheme

Identifying nature and causes of errors

source:
Please come back to us if you need any further assistance.
translation:
Bitte **{kommen Sie zu uns zurück}**/*style+awkward*, wenn Sie weitere Hilfe benötigen.

- Error marking: words in bold are problematic
- Error classification: words in bold are problematic because the style is awkward

? But: why/how did this error happen at all?

! the text was translated word by word from English without necessary rephrasing

⇒ cause/nature of this issue:
rephrasing, *literal translation*

Aggregating scores from annotations

From segment scores to total scores

the value for one entire text is obtained by aggregating values of all its sentences

aggregating segment-level values

- ranking:
  a common idea = compute the average number of comparisons where each system was judged better than other system
- scoring:
  average value over all sentences

From error annotations to scores

This is a **[totally silly text]**, just for **[example]**. It is not even translated from **[any language]**. It only serves to illustrate **[]** error counts.

extracting scores from error annotations

- word-level error count: 7
- word-level error rate: 7/29 = 24.1%
- span-level error count: 4
- different weights for major and minor errors make sense in any case

From error annotations to scores

This is a **[totally silly text]**, just for **[example]**. It is not even translated from [any language]. It only serves to illustrate **[]** error counts.

different weights for major and minor errors

for example: major errors = 1, minor errors = 0.1

- word-level error count: 5 major + 2 minor = 5.2
- word-level error rate: 5.2/29 = 17.9%
- span-level error count: 3 major + 1 minor = 3.1

Subjectivity

Different evaluators = different results

- different evaluators will
  assign different scores / mark different error spans / assign different error labels
- even the same evaluator at different times

There is no single correct translation of a text

- different evaluators might have in mind different correct translations
- errors might be perceived differently and annotated differently
- people have personal preferences

Measuring subjectivity

inter-annotator agreement (IAA)

- measuring similarity between the annotations of different evaluators
- several methods to do that
- the details depend on the exact type and process of evaluation

intra-annotator agreement

- measuring similarity between the annotations of the same evaluator at different times

two annotators: adequacy and fluency

		adequacy		fluency	
translation	text	e1	e2	e1	e2
reference	It will be a sort of bridge.	5	5	5	4
system 1	It is almost as a bridge act.	2	3	1	2
system 2	It will act as a bridge.	5	4	5	5
system 3	It will not act as a bridge.	1	2	5	5
system 4	It will sort of bridge be.	4	5	2	4

two annotators: error marking

system		errors	err. count	error overlap
reference	e1	It will be a sort of bridge.	0	no errors
	e2	It will be a sort of bridge.	0	
system 1	e1	It is almost **as** a bridge **act** .	2	1/2
	e2	It **is** almost as a bridge **act** .	2	(*act*/*is*, *act*)
system 2	e1	It will act as a bridge.	0	no errors
	e2	It will act as a bridge.	0	
system 3	e1	It will **not** act as a bridge.	1	1/3
	e2	It will **not act as** a bridge.	3	(*not*/*not*, *act*, *as*)
system 4	e1	It will **[]** sort of bridge **be** .	2	1/4
	e2	It will **sort of** bridge **be** .	3	(*be*/*[]*, *be*, *sort*, *of*)

two annotators: error classification

not only error spans but also error types

It is almost as a bridge act.

- “is” should be “will” → mistranslation
- “almost” = addition
- “act” is in the wrong position → order

or

- “will” is missing → omission
- “is” = incorrect form of “be” → tense
- “almost as” should be ”sort of” → mistranslation
- “act” = addition

All in all

human evaluation is

- important
- requires knowledge
- requires time
- intrinsically subjective
  (subjectivity can be reduced but not eliminated)

- human evaluation should be carried out whenever possible
- often it is not (developing a MT system)

$\Rightarrow$ automatic evaluation

Automatic evaluation

Automatic evaluation

automatic evaluation = approximation of human evaluation

it has to overcome the disadvantages of manual evaluation

- it has to be fast
- it has to be cheap (not requiring large resources)
- it has to be consistent (always generating the same result)

it has to be correct

what does that mean?

Evaluating automatic evaluation metrics

- they certainly are fast, cheap and consistent
- but how correct are they?

compare the automatic metric scores with manual evaluation scores

correlation coefficients:

- around 0 $\rightarrow$ the metric is not good at all
  practically no similarity between the human and automatic scores
- close to 1 $\rightarrow$ the metric perfectly substitutes the manual score
- close to -1 $\rightarrow$ also a “perfect” metric, only reversed

Types of automatic metrics

based on overlap between the MT output and a reference human translation

- similarity (n-gram matching)
  how similar is the MT output to a reference human translation
- dissimilarity (edit distance)
  how different is the MT output from a reference human translation

based on neural networks

- a neural network trained to generate a score

Inputs of an automatic metric

obviously, the translation to be evaluated
what else?

- for overlap-based metrics
  - a reference human translation

- for neural-based metrics
  many possiblities
  - a reference human translation
  - and/or other correct translations
  - and/or human scores
  - and/or the source text
  - if without a reference translation: ”quality estimation”

Overlap-based metrics

Similarity (matching)

precision and recall

- Precision shows how many instances in the *generated set* can be found in the *reference set *(and therefore considered as correct).

$$precision = \frac{N(matches\_in\_Translation\_Output)}{N(instances\_in\_Translation\_Output)}$$

- Recall shows how many instances in the *reference set* are covered by the *generated set*.

$$recall = \frac{N(matches\_in\_Reference\_Translation)}{N(instances\_in\_Reference\_Translation)}$$

What are 'instances'?

- usually sequences of *n* words (*n*-grams)
- if *n*=1 (word matching):

reference	**It will be** a **sort of bridge .**
system 4	**It will sort of bridge be .**
precision	7/7 (100%)
recall	7/8 (87.5%)

punctuation is considered as a separate word

Word order matters, too

- checking only word matches does not take word order into account
- checking for matching groups of words does
- if $n = 2$ (bigrams)

reference	**It**∼**will** will∼be be∼a a∼sort **sort**∼**of of**∼**bridge** bridge∼.
system 4	**It**∼**will** will∼sort **sort**∼**of of**∼**bridge** bridge∼be be∼.
precision	3/6 (50.0%)
recall	3/7 (42.8%)

precision and recall much lower than for single words

Trigrams ($n = 3$) and fourgrams ($n = 4$)

reference	It∼will∼be will∼be∼a a∼sort∼of **sort∼of∼bridge** of∼bridge∼.
system 4	It∼will∼sort will∼sort∼of **sort∼of∼bridge** of∼bridge∼be bridge∼be∼.
precision	1/5 (20.0%)
recall	1/6 (16.7%)

only one single trigram match

4-gram: no matches (precision = 0, recall = 0)

Joining all *n*-gram matches

Arithmetic mean:

$$\textit{precision} = \frac{100 + 50 + 20 + 0}{4} = 42.5$$

$$\textit{recall} = \frac{87.5 + 42.8 + 16.7 + 0}{4} = 36.75$$

Geometric mean:

$$\textit{precision} = \sqrt[4]{100 \cdot 50 \cdot 20 \cdot 0} = 0$$

$$\textit{recall} = \sqrt[4]{87.5 \cdot 42.8 \cdot 16.7 \cdot 0} = 0$$

Joining precision and recall: F-score

$$F_\beta = (1 + \beta^2)\frac{2 \cdot \mathit{precision} \cdot \mathit{recall}}{\beta^2 \cdot \mathit{precision} + \mathit{recall}}$$

- $\beta = 1 \Rightarrow$ simple harmonic mean (also called F1)
- $\beta > 1 \Rightarrow$ recall has more weight
- $\beta < 1 \Rightarrow$ precision has more weight

n-gram matching metrics

- BLEU
  only precision, 'brevity penalty' instead of recall
  geometric mean of word 4-grams

- METEOR
  both precision and recall, more weight to recall
  accepts common word stems, synonyms, paraphrases

- chrF
  both precision and recall, more weight to recall
  arithmetic mean of character 6-grams (+ word bigrams for chrF++)
  arithmetic mean

Dissimilarity (distance)

edit (Levensthein) distance

the minimum number of corrections (“edits”) necessary to transform the translation output into the reference translation

$$N(\textit{substitutions}) + N(\textit{deletions}) + N(\textit{insertions})$$

- **Substitution**: one word is replaced by another
- **Deletion**: a word is missing and should be added
- **Insertion**: a word is inserted and should be removed

Edit distance and alignment: examples

reference:	It will	be$_{del.}$	a$_{del.}$	sort	of	bridge		.
system 4:	It will			sort	of	bridge	be$_{ins.}$	.

two deletions, one insertion, no substitutions
$\Rightarrow$ edit distance = 3

reference:	It will	be$_{subst.}$	a$_{subst.}$	sort$_{subst.}$	of$_{del.}$	bridge	.
system 2:	It will	act$_{subst.}$	as$_{subst.}$	a$_{subst.}$		bridge	.

one deletion, three substitutions
$\Rightarrow$ edit distance = 4

Edit distance and different alignments

- edit distance does not depend on the exact implementation (which operation is first checked in the code)
- alignment does

Different alignments for system 2

reference:	It will	be$_{subst.}$	a$_{subst.}$	sort$_{subst.}$	of$_{del.}$	bridge	.
system 2:	It will	act$_{subst.}$	as$_{subst.}$	a$_{subst.}$		bridge	.

one deletion, three substitutions $\Rightarrow$ edit distance = 4

reference:	It will	be$_{del.}$	a$_{subst.}$	sort$_{subst.}$	of$_{subst.}$	bridge	.
system 2:	It will		act$_{subst.}$	as$_{subst.}$	a$_{subst.}$	bridge	.

one deletion, three substitutions, only the particular words are tagged differently

reference:	It will	be$_{subst.}$		a	sort$_{del.}$	of$_{del.}$	bridge	.
system 2:	It will	act$_{subst.}$	as$_{ins.}$	a			bridge	.

one substitution, two deletions, one insertion $\Rightarrow$ edit distance = 4

Edit distance metrics

- Word Error Rate (WER) – edit distance (or Levenshtein distance) itself normalised by the length of the reference translation

$$WER = \frac{N(substitutions) + N(deletions) + N(insertions)}{reference\_length}$$

- Translation Edit Rate (TER) – edit distance extended by ”block shift” operation (move a word or group of words (’block’) into another position)

$$TER = \frac{N(substitutions) + N(deletions) + N(insertions) + N(block\_shifts)}{reference\_length}$$

WER vs TER

reference:	It will	be$_{del.}$	a$_{del.}$	sort	of	bridge		.
system 4:	It will			sort	of	bridge	be$_{insertion}$	.

WER = 3/8 = 37.5%

reference:	It will	be$_{shift}$	a$_{del.}$	sort	of	bridge		.
system 4:	It will			sort	of	bridge	be$_{shift}$	.

TER = 2/8 = 25.0%.

WER vs TER

ref:	It will	be$_{del.}$ considered$_{del.}$ as$_{del.}$ a$_{del.}$	sort of bridge		.
sys 4^{+}:	It will		sort of bridge	be$_{ins.}$ considered$_{ins.}$ as$_{ins.}$	.

4 deletions, 3 insertions
edit distance = 7, WER = 7/10 = 70%

ref:	It will	[be considered as]$_{shift}$ a$_{del.}$	sort of bridge		.
sys 4^{+}:	It will		sort of bridge	[be considered as]$_{shift}$	.

“be considered as” = block (a group or chunk) of successive words in the wrong place
TER = 2/8 = 25.0%

A brief history of the common overlap-based metrics

- WER: the very first metric in MT, inherited from speech recognition
  but: not appropriate for MT
  generally too harsh as it penalises a number of acceptable variations
- BLEU (2002): the very first metric designed for MT
  + old habits die hard so it is still used
- METEOR (2005): using recall, too, and allowing synonyms, stems and paraphrases
- TER (2006): adapting WER to MT by introducing block shifts
- chrF (2015): character *n*-grams instead of word *n*-grams
  the best overlap-based metric

Overlap-based metrics: properties

+ the obtained score is explicable
  it reflects a degree of (dis)similarity between the two texts
− but not interpretable
  it does not give any information about any translation quality criterion
− penalise correct translations different from the given reference translation
− do not correlate very well with human scores

+ language-independent
+ low computational cost
+ fairly reproducible
  (variances possible due to exact implementation and performed tokenisation)

Neural metrics

Current trend: neural metrics

current research effort in automatic MT evaluation is almost exclusively focusing on neural metrics

the main advantages of neural metrics

- not just a simple overlap between two texts
- can reward correct translations different from the given reference translation
- better correlation with human scores

Neural metrics: disadvantages

- − score unexplainable (why exactly this value?)
- − difficult to reproduce (a large number of hyper-parameters)
- − language-dependent
- − high computational cost
- − difficult to maintain
  (external changes in PyTorch, CUDA, etc.)
- ± require training
  - + advantage if specific training data is available
  - − disadvantage if it is not

Some important neural metrics

- BERTscore
  - cosine similarity between BERT word representations in the evaluated sentence and those in the reference
- COMET
  - XLM-RoBERTa fine-tuned on human scores
  - takes source text, too
- BLEURT
  - BERT model fine-tuned in two phrases
  - first phase on synthetic data, second phase on human scores
- PRISM
  - considers MT output as a paraphrase of the reference
  - trained on a large multilingual parallel data through a multilingual NMT framework
  - does not need human scores for training

All in all

- neural metrics better correlate with human scores
- overlap metrics still widely used (often more simple and easier to run)

recommendations

- neural metrics:
  report scores of at least one neural metric if your language pair is covered
- overlap metrics:
  report chrF scores (not BLEU scores)
- perform at least a small-scale simple human evaluation

Test Suites

What are test suites (or challenge test sets)?

test sets specialised for evaluation of particular (usually linguistic) aspects/phenomena, such as

- ambiguous words
- negation
- named entities
- gender
- ...

Why to use such test sets?

the traditional “standard” test sets

+ reflect a “real world situation”
– have an uneven distribution of specific (linguistic or other) phenomena

challenge test sets

+ have a good coverage of the phenomenon/phenomena of interest
– do not reflect the “real world”, namely distribution of phenomena in naturally ocurring texts (news articles, user comments, conversations, etc.)

A brief history of test suites

- first test suites designed to evaluate rule-based systems
  (how they are dealing with syntax and morphology)
- emergency of SMT suppressed the usage of test suites
- emergency of NMT revived the usage of test suites
  modern test suites do not cover only syntax and morphology, but a wide range of different aspects

Example of a standard test set

1) 'The sea is ours': landlocked Bolivia hopes court will reopen path to Pacific.
2) Naval bases from Lake Titicaca to the Amazon are daubed with the motto: "The sea is ours by right.
3) To recover it is a duty."
4) Throughout landlocked Bolivia, the memory of a coastline lost to Chile in a bloody 19th-century resource conflict is still vivid - as is the yearning to sail the Pacific Ocean once more.
5) Those hopes are perhaps at their highest in decades, as Bolivia awaits a ruling by the international court of justice on 1 October after five years of deliberations.
6) "Bolivia has the momentum, a spirit of unity and serenity, and is of course expecting with a positive view the outcome," said Roberto Calzadilla, a Bolivian diplomat.
7) Many Bolivians will watch the ICJ ruling on big screens across the country, hopeful that the tribunal in The Hague will find in favour of Bolivia's claim that - after decades of fitful talks - Chile is obliged to negotiate granting Bolivia a sovereign outlet to the sea.
8) Evo Morales, Bolivia's charismatic indigenous president - who faces a controversial battle for re-election next year - also has plenty riding on Monday's ruling.
9) "We are very close to returning to the Pacific Ocean," he vowed in late August.
10) But some analysts believe that the court is unlikely to decide in Bolivia's favour - and that little would change if it did.
11) The Netherlands-based UN body has no power to award Chilean territory, and has stipulated that it will not determine the outcome of possible talks.
12) And far from furthering Bolivia's cause, the past four years may have set it back.
13) Bolivia and Chile will at some point continue to talk, but it will be extremely difficult to hold discussions after this.

What about pronoun “it” and its referent?

1) ’The sea is ours’: landlocked Bolivia hopes court will reopen path to Pacific.
2) Naval bases from Lake Titicaca to the Amazon are daubed with the motto: "The **sea** is ours by right.
3) To recover **it** is a duty."
4) Throughout landlocked Bolivia, the memory of a coastline lost to Chile in a bloody 19th-century resource conflict is still vivid - as is the yearning to sail the Pacific Ocean once more.
5) Those hopes are perhaps at their highest in decades, as Bolivia awaits a ruling by the international court of justice on 1 October after five years of deliberations.
6) "Bolivia has the momentum, a spirit of unity and serenity, and is of course expecting with a positive view the outcome," said Roberto Calzadilla, a Bolivian diplomat.
7) Many Bolivians will watch the ICJ ruling on big screens across the country, hopeful that the tribunal in The Hague will find in favour of Bolivia’s claim that - after decades of fitful talks - Chile is obliged to negotiate granting Bolivia a sovereign outlet to the sea.
8) Evo Morales, Bolivia’s charismatic indigenous president - who faces a controversial battle for re-election next year - also has plenty riding on Monday’s ruling.
9) "We are very close to returning to the Pacific Ocean," he vowed in late August.
10) But some analysts believe that **the court** is unlikely to decide in Bolivia’s favour - and that little would change if **it** did.
11) The Netherlands-based UN body has no power to award Chilean territory, and has stipulated that **it** will not determine the outcome of possible talks.
12) And far from furthering **Bolivia’s cause**, the past four years may have set **it** back.
13) Bolivia and Chile will at some point continue to talk, but **it** will be extremely difficult to hold discussions after this.

What about pronoun “it” and its referent?

- pronoun “it” appears only in 5 segments out of 13
- of those 5 “its”, only 3 are really referential (2)+3), 10), and 12)

→ a potential test suite for “it” and its referent:

Naval bases from Lake Titicaca to the Amazon are daubed with the motto: ”The **sea** is ours by right. To recover **it** is a duty.”
But some analysts believe that **the court** is unlikely to decide in Bolivia’s favour - and that little would change if **it** did.
And far from furthering **Bolivia’s cause**, the past four years may have set **it** back.

What about this test set?

Oh, no, there's another **fly**, I hate them! Don't worry, I'll kill **it** for you.
How about a **sandwich**? I only have one, but we could share **it**...
This is a very good **chair**. It really is, I bought it yesterday.
Their **juice** was delicious. I'm wondering why don't they make **it** anymore?
It's been a while since I last went to the **river**. It feels great to see **it** again.
So you took his **bag**? Yes, he left **it** at home.
I think you should change your **hair colour**. If you don't like **it**, don't look at **it**.
You ruined my **jacket**. **It** was already ruined.
Here's your **hat**. Thank you for bringing **it** to me.
This **bike** is so ugly. Why do you like **it** so much?
Where is your **cat**? **It** is in the living room.
Where is your **dog**? **It** is in the kitchen.

How are challenge test sets evaluated?

- only the phenomenon of interest is evaluated
- other errors are not considered
- accuracy:
  percentage of entries where the phenomenon is correctly translated

Any questions?