Unsupervised machine translation and repair

Colt Dahl

Outline

1. Introduction
2. Supervised machine translation
3. Unsupervised machine translation
4. TransRepair
5. Results

Introduction

The benefits of machine translation:

- Cheaper
- Time-effective
- No need for independent human translator

Introduction

The benefits of machine translation:

- Cheaper
- Time-effective
- No need for independent human translator

The drawbacks of machine translation:

- Quality of translation
- Lost subtleties, ambiguity
- Misunderstanding of context

Rule-based machine translation

- Grammar and vocabulary retrieved from dictionaries written by linguists/technicians
- Rules represent language structure
- Rules are mapped between languages to produce translations for each language pair (i.e. English->French)

Rule-based machine translation

- Grammar and vocabulary retrieved from dictionaries written by linguists/technicians
- Rules represent language structure
- Rules are mapped between languages to produce translations for each language pair (i.e. English->French)

- Known as the “classical approach”
- List of rules becomes unwieldy fast, making it harder for distant languages

Rule-based machine translation

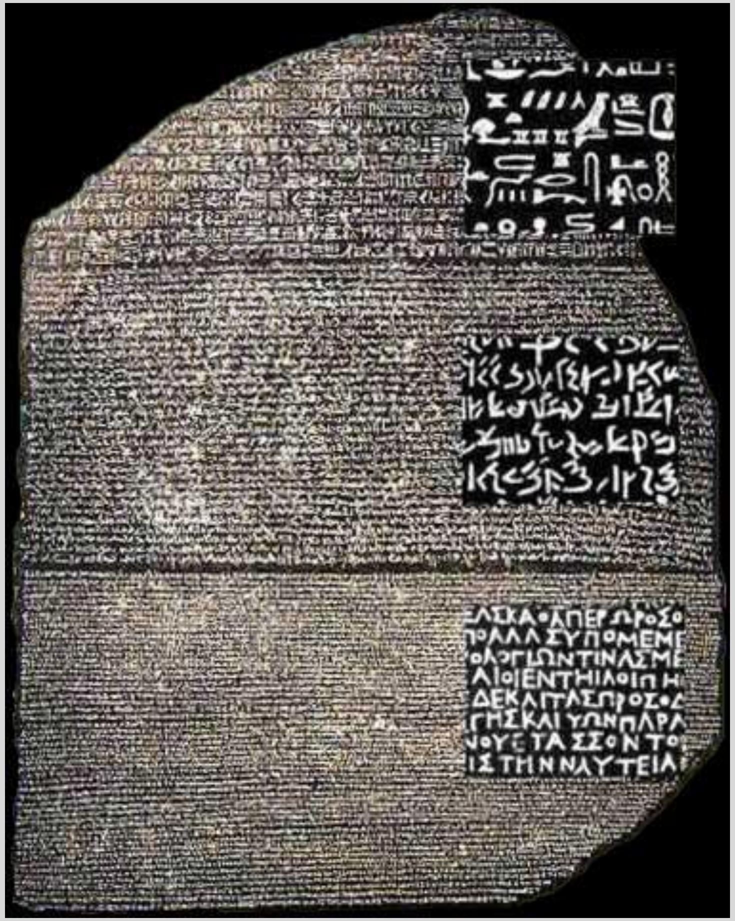
- Grammar and vocabulary retrieved from dictionaries written by linguists/technicians
- Rules represent language structure
- Rules are mapped between languages to produce translations for each language pair (i.e. English->French)

- Known as the “classical approach”
- List of rules becomes unwieldy fast, making it harder for distant languages

Can we think of a way of making the induction of these rules easier?

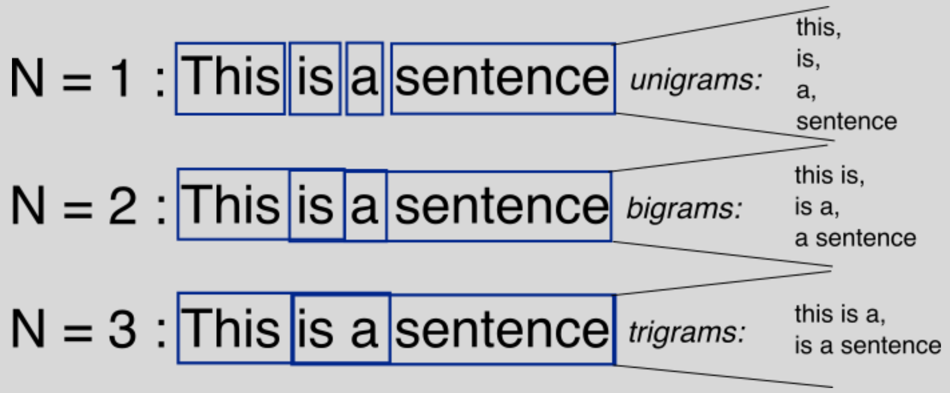
Statistical machine translation (SMT)

- Based on analysis of "parallel texts" most often as "bilingual corpora", a structured set of texts between two languages
- Attempts to match source language phrases with the most similar phrases in target language
- The Rosetta stone, discovered in 1799, is a parallel text



Statistical machine translation (SMT)

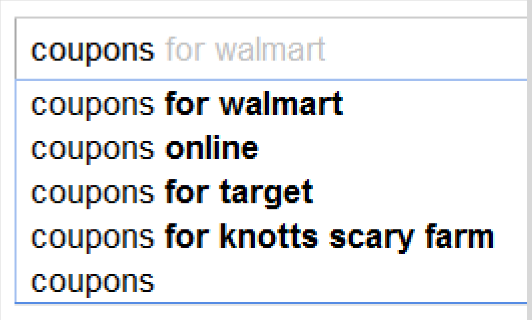
- Based on analysis of "parallel texts" known as bilingual corpora, structured set of texts
- The Rosetta stone, discovered in 1799, is a parallel text
- Attempts to match source language phrases with the most similar phrases in target language
- Isolates pieces of sentences through use of "n-grams"
- Makes predictions on most probable output



N-grams

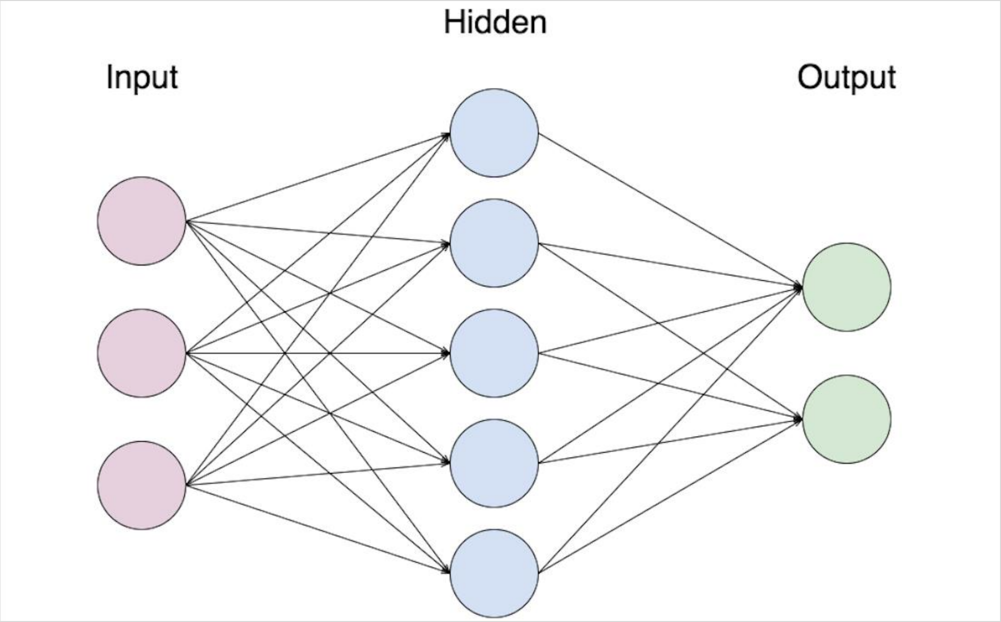
Statistical machine translation (SMT)

- Examples of predictive output through statistical methods
- Your phone's autocomplete function
- Search engine term suggestions



Neural machine translation (NMT)

- Takes source parallel texts, like statistical machine translation
- Inputs are sent through hidden nodes
- If MT output does not match good quality translation, biases and weights for nodes/edges are changed
- Effective through training
- Quickly surpassed every other translation method



Drawbacks to bilingual corpora

- Bilingual corpora is expensive to maintain and produce
- Why don’t we try to train machine translators on monolingual corpora?

Drawbacks to bilingual corpora

- Bilingual corpora is expensive to maintain and produce
- Why don't we try to train machine translators on monolingual corpora?

Unsupervised Machine Translation

- Machine translation predominantly using synthetic corpora, instead of exclusively manually produced corpora.

Iterative training of unsupervised machine translation

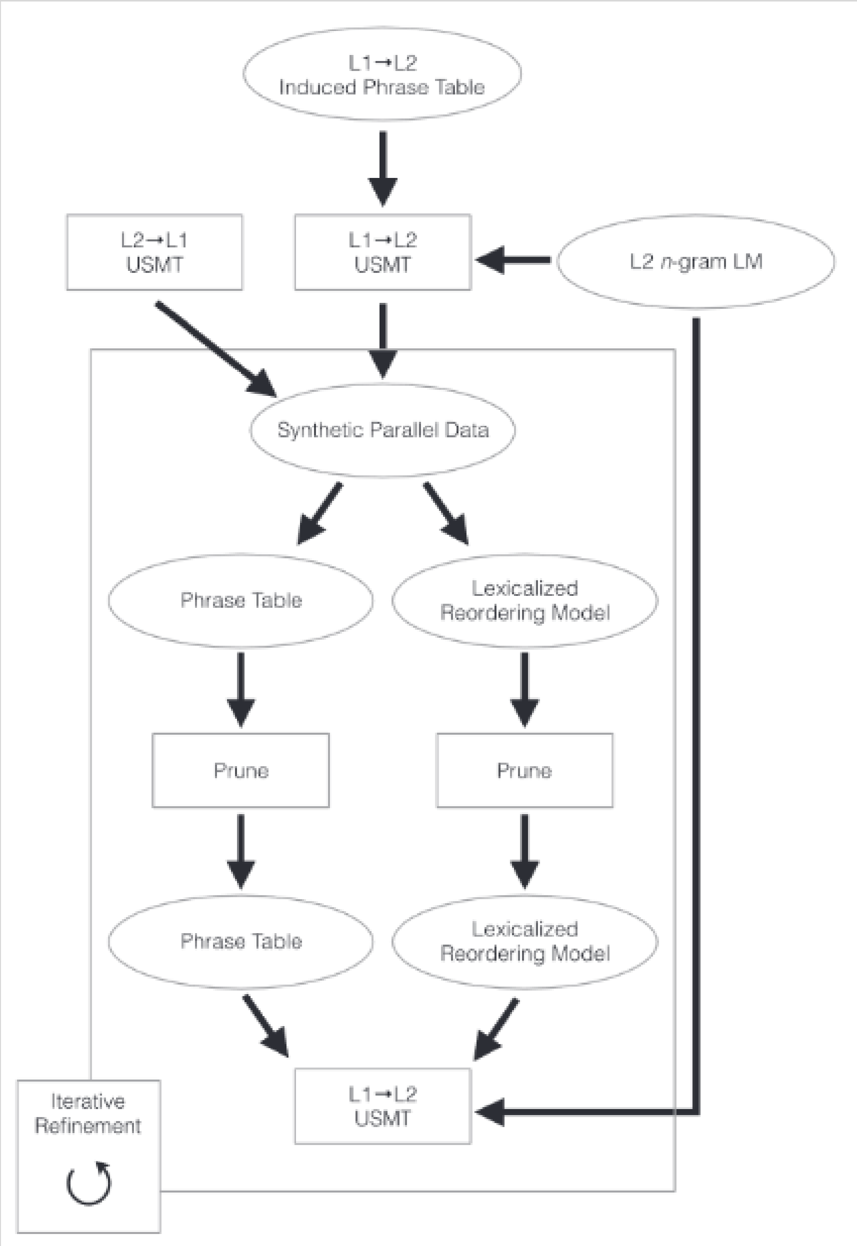
- We will be using a combination of the two previous systems to perform unsupervised machine translation and induce synthetic parallel data
- Unsupervised statistical machine translation (USMT) and unsupervised neural machine translation (UNMT) iterating on each other.

Unsupervised statistical machine translation

◦ Unsupervised MT uses a lot less manually produced parallel data for training.

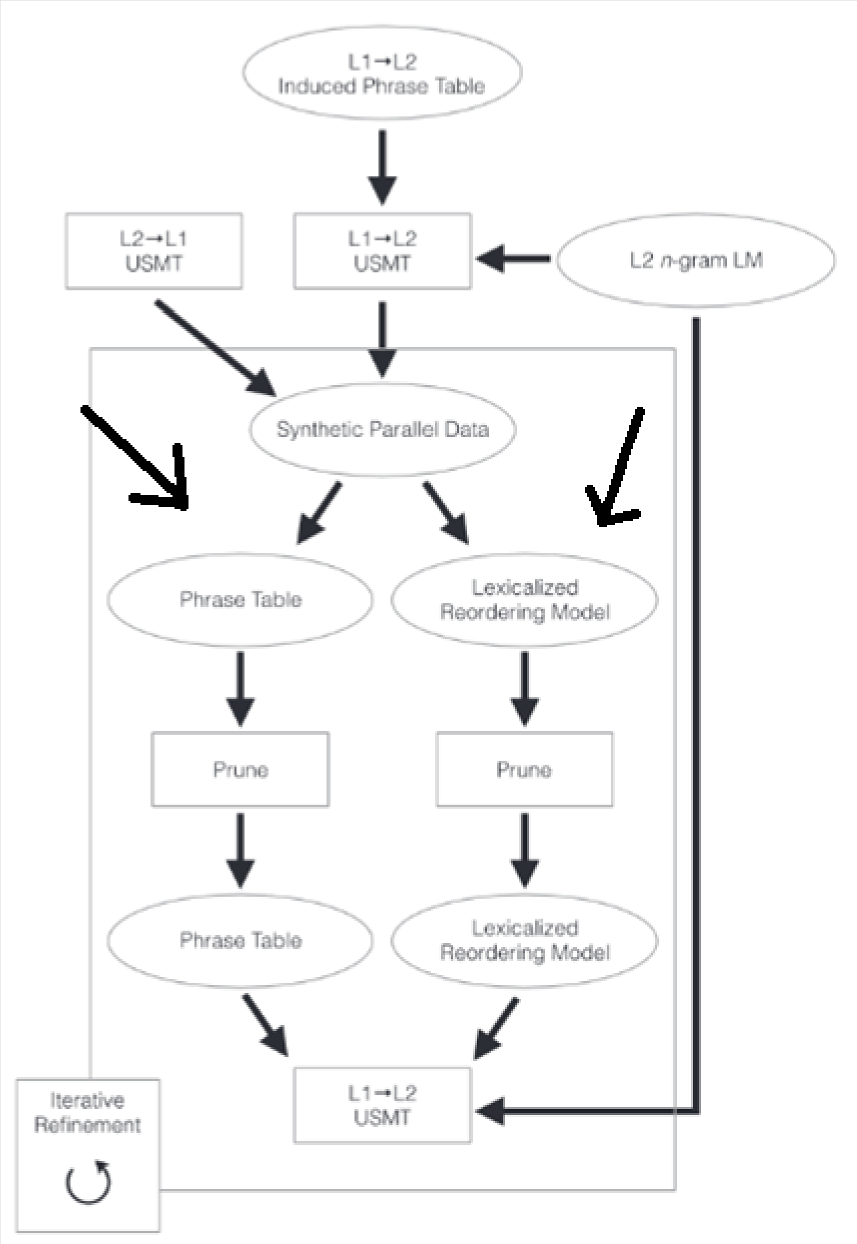
Phrase table induction

The induction of synthetic corpora.



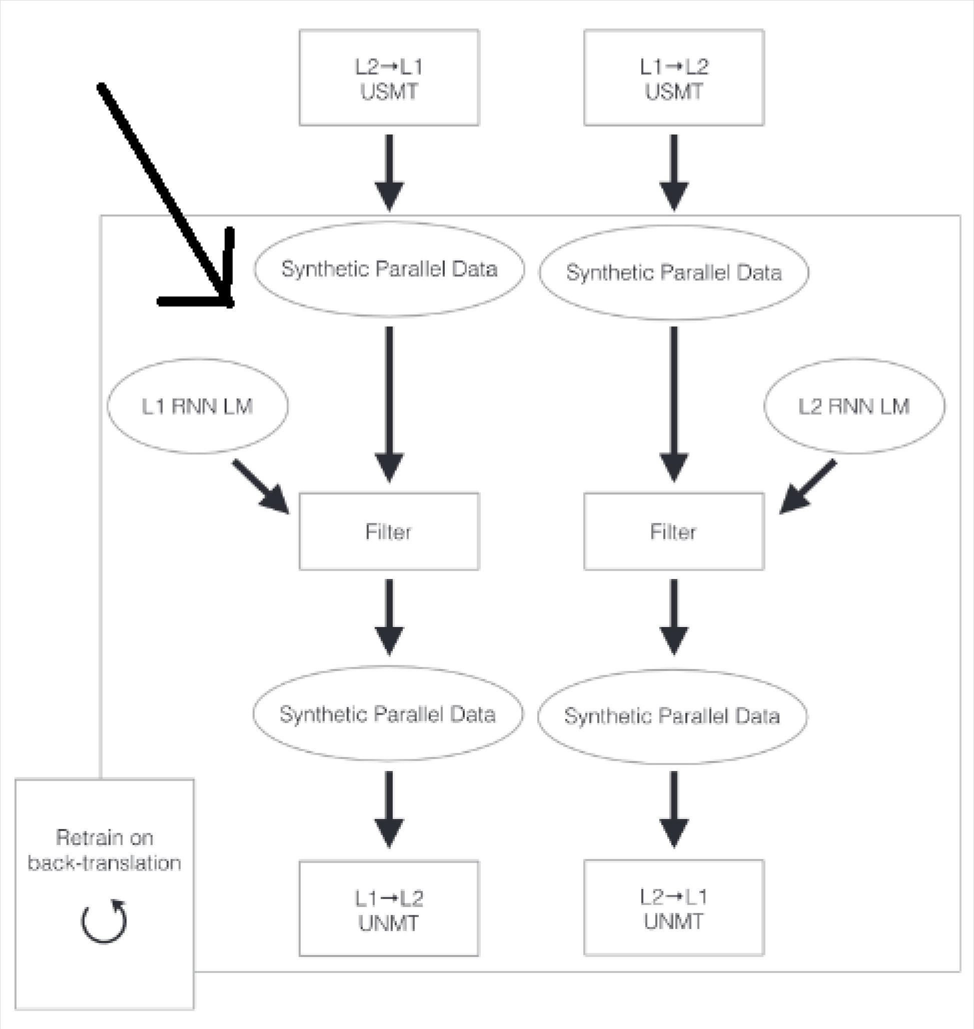
Pruning and the Lexical reordering model

Phrase tables and the Lexicalized reordering model go through pruning, and then the translation is refined through the USMT system again.



Unsupervised neural machine translation

Assuming that USMT produces translations of a reasonable quality, we first train L1→L2 and L2→L1 UNMT systems on sentence pairs respectively generated by translating data with L2→L1 and L1→L2 USMT systems.



BLEU scores

- BiLingual Evaluation Understudy
- Common method machine translation precision is measured
- The BLEU score is an algorithm, comparing the translations of the machine's output from a phrase's input to that of a good-quality translation
- Most commonly measured 0-100 for ease of communication

BiLingual Evaluation Understudy (BLEU)

French: "Le chat est sur le tapis.“ <- Sentence for translation.
Reference 1: "The cat is on the mat.“ <- Good quality manual translation.
Reference 2: "There is a cat on the mat.“ <-Good quality manual translation.
Candidate: "The the the the the the the" <-Machine translation output.

Looks at each word in the candidate phrase and looks to see if it appears in the references, and if so, it gets a point.

Unmodified Precision: 7/7

BiLingual Evaluation Understudy (BLEU)

French: "Le chat est sur le tapis.“ <- Sentence for translation.
Reference 1: "The cat is on the mat.“ <- Good quality manual translation.
Reference 2: "There is a cat on the mat.“ <-Good quality manual translation.
Candidate: "The the the the the the the" <-Machine translation output.

Looks at each word in the candidate phrase and looks to see if it appears in the references, and if so, it gets a point.

Unmodified Precision: 7/7 -> Modified precision: 2/7

Modified precision sets the max number of times an isolate will be counted, since the max number of times we say “the” in either reference is now 2, the candidate sentence now has a precision of 2/7.

BiLingual Evaluation Understudy (BLEU)

French: "Le chat est sur le tapis.“ <- Sentence for translation.
Reference 1: "The cat is on the mat.“ <- Good quality manual translation.
Reference 2: "There is a cat on the mat.“ <-Good quality manual translation.
Candidate: "The the the the the the the." <-Machine translation output.

Looks at each n-gram in the candidate phrase to see if it appears in the references, and if so, it gets a point.

Unmodified Precision: 7/7 -> Modified precision: 2/7

Modified precision sets the max number of times an isolate will be counted, since the max number of times we say “the” in either reference is now 2, the candidate sentence now has a precision of 2/7.

To improve it even further, we increase the n-gram value from 1 to 2 and begin to receive a modified precision of 0, as “the the” appears nowhere in either of the good quality references.

A candidate such as “The cat is on mat” with the use of higher n-grams would then receive a better score than “The the the the the the”.

Results

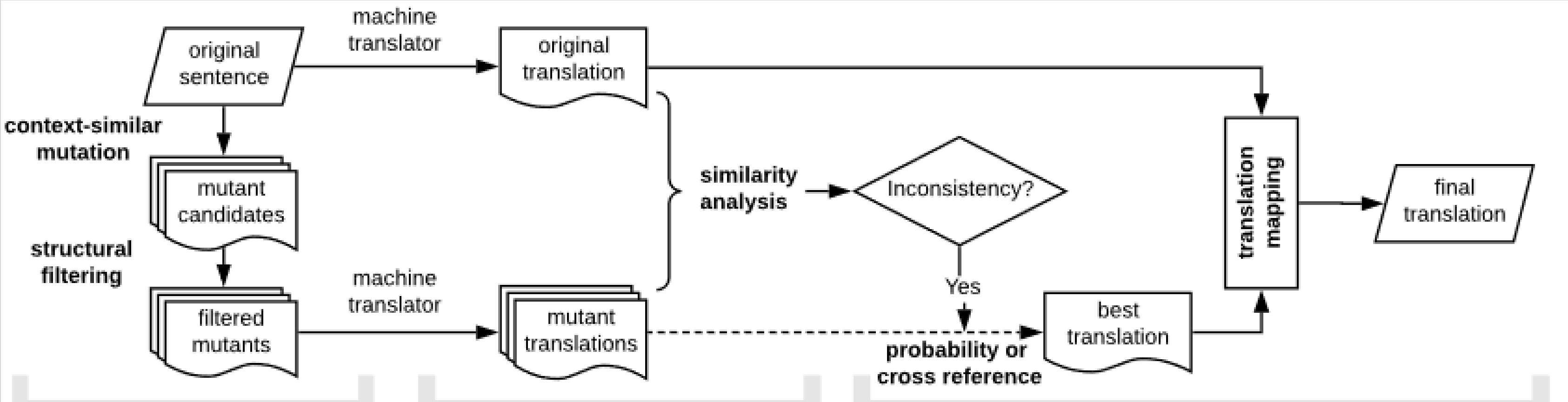
Iterative training of USMT/UNMT

Received a BLEU score, on average, a few points higher than previous methods for unsupervised machine translation.

System	Newstest				NTCIR		
	de→en	en→de	fr→en	en→fr	ja→en	en→ja	#
Lample et al. [24]'s USMT	22.1	17.5	26.2	23.9	20.5	21.6	1
Lample et al. [24]'s UNMT	20.3	17.0	23.6	22.9	15.8	17.2	2
USMT-1	23.4	18.8	26.7	25.3	21.3	22.0	3
↳ UNMT-1	29.4	22.8	28.8	28.1	25.3	27.8	4
↳ USMT-2	26.6	21.4	28.0	27.3	21.6	25.0	5
↳ UNMT-2	**30.4**	**24.3**	**29.2**	**29.0**	25.9	**29.2**	6
UNMT-1 ($P = 8.5 \times 10^6$)	29.8	22.8	28.9	28.4	**26.0**	28.0	7
Supervised SMT	30.4	26.4	35.3	32.7	27.6	31.3	8
Supervised NMT	35.8	32.9	35.9	37.2	43.5	48.7	9

TransRepair

- Takes a sentence as input and applies a contextual mutation
- Compares the original with the mutated without hurting the structure of the sentence
- Similarity analysis is done
- If an inconsistency is found, it is repaired in line with the mutant sentences



TransRepair

- Takes a sentence as input and applies a contextual mutation
- Compares the original with the mutated without hurting the structure of the sentence
- Similarity analysis is done
- If an inconsistency is found, it is repaired in line with the mutant sentences

Input	**Original translation**	**Repaired translation**
Female students do **good** research in computer science.	nüxuesheng zai jisuanji kexue fangmian zuole **hen-duo** yanjiu *[Bug: "good" → "a lot".]*	nüxueshengzai jisuanji kexue fangmian zuole **hen-haode** yanjiu
If you need help, you can enjoy timely services by press-ing a nearby **one of** the 41 call buttons in the station.	ruguo ni xuyao bangzhu, ni keyi tongguo an fujin de 41 ge hujiao anniu xiangshou jishi de fuwu. *[Bug: "one of" is not translated.]*	ruguo ni xuyao bangzhu, ni keyi tongguo an fujin de 41 ge hujiao anniu **zhong de yige** lai xiangshou jishi de fuwu.

Results

Translation improvement with TransRepair (manual inspection)

	Aspect	Improved	Unchanged	Decreased
GT.LCS	**Translation consistency**	33 (85%)	4 (10%)	2 (5%)
	Translation acceptability: overall	22 (28%)	48 (62%)	8 (10%)
	Translation acceptability: original	10 (26%)	23 (59%)	6 (15%)
	Translation acceptability: mutant	12 (31%)	25 (64%)	2 (5%)
Trans.Prob	**Translation consistency**	51 (88%)	6 (10%)	1 (2%)
	Translation acceptability: overall	30 (26%)	76 (66%)	10 (9%)
	Translation acceptability: original	15 (26%)	36 (62%)	7 (12%)
	Translation acceptability: mutant	15 (26%)	40 (69%)	3 (5%)

Questions

Acknowledgements

- https://dl.acm.org/doi/10.1145/3389790 - Iterative training of Unsupervised neural and statistical machine translation, diagrams, tables
- https://dl.acm.org/doi/pdf/10.1145/3377811.3380420 - TransRepair, diagrams, tables
- https://searchengineland.com/how-google-instant-autocomplete-suggestions-work-62592 - Image
- https://discoveringegypt.com/egyptian-video-documentaries - Image
- https://towardsdatascience.com/classical-neural-network-what-really-are-nodes-and-layers-ec51c6122e09 - Image