Comparing Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) Performances

Hervé Blanchon
Laurent Besacier
Laboratoire LIG
Équipe GETALP
herve.blanchon@univ-grenoble-alpes.fr
laurent.besacier@univ-grenoble-alpes.fr

Outline

- Introduction
  - SMT and NMT in a Nutshell
- Thesis
  - NMT is great – paper #1
- Antithesis
  - NMT is not so great, sometimes SMT wins – paper #2
- Synthesis:
  - NMT is promising to tackle hard challenges – paper #3

SMT and NMT in a nutshell

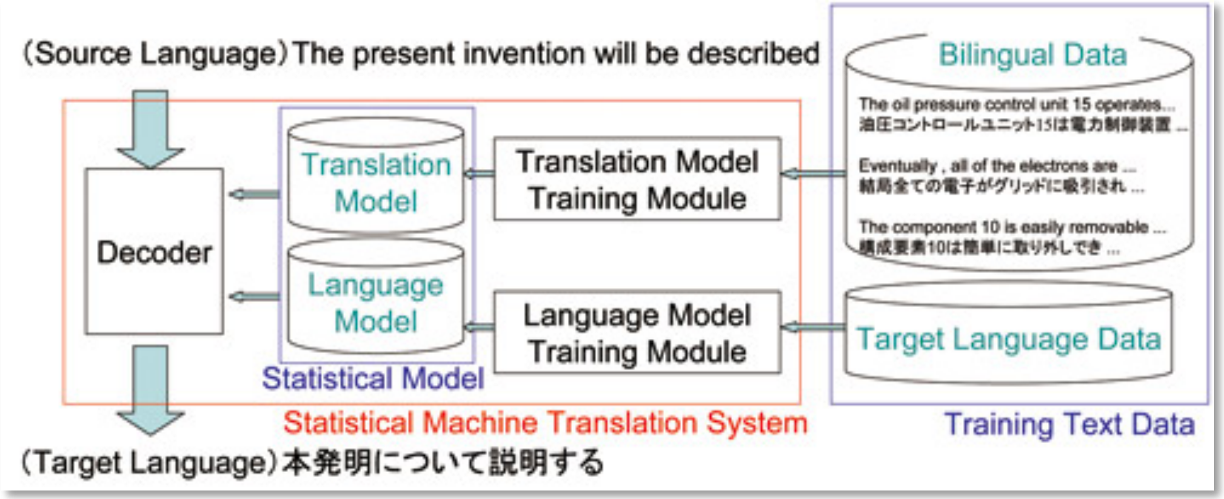
INTRODUCTION

Statistical Machine Translation (SMT)

- Built on pioneering work at IBM in the 1990s
  - P. Brown & al. *The mathematics of statistical machine translation: parameter estimation* (1993)
  - Bayesian framework, formalized word alignment concept, etc.
- Models later extended to phrases
  - P. Koehn & al. *Statistical phrase-based translation* (2003)
    - Lead to **Moses** open source toolkit in 2007
  - Largely used in academia and industry since then

Statistical Machine Translation (SMT)

- Overview[1]



- Key component: phrase table[2]

```
" التطورات ||| " developments ||| 1 0.506684 1 0.863476 2.718 ||| ||| 7 7
" التطورات في ||| " developments in the ||| 1 0.336365 0.461538 0.129703 2.718 ||| ||| 6 13
" التطورات في ||| " developments in ||| 1 0.336365 0.538462 0.349129 2.718 ||| ||| 7 13
" التطورات في ميدان ||| " developments in the field of ||| 0.857143 0.170757 0.5 0.017437 2.718 ||| ||| 7 12
" التطورات في ميدان ||| " developments in the field ||| 0.857143 0.170757 0.5 0.0880919 2.718 ||| ||| 7 12
" التطورات في ميدان المعلومات ||| " developments in the field of information ||| 0.857143 0.0726128 1 0.0158491 2.718 ||| ||| 7 6
" التطورات في ميدان ||| " developments in the field of ||| 0.142857 0.000431254 0.5 0.0172767 2.718 ||| ||| 7 2
" التطورات في ميدان ||| " developments in the field ||| 0.142857 0.000431254 0.5 0.0872822 2.718 ||| ||| 7 2
" التطورات في ميدان المعلومات ||| " developments in the field of information ||| 0.142857 0.000183387 1 0.0157034 2.718 ||| ||| 7 1
" التعاون ||| " cooperation ||| 0.965517 0.619446 0.982456 0.894784 2.718 ||| ||| 58 57
" التعاون ||| " enhanced ||| 1 0.0100982 0.0175439 0.00079395 2.718 ||| ||| 1 57
```

Credits:

1 http://www.kecl.ntt.co.jp/rps/_src/sc1134/innovative_3_1e.jpg

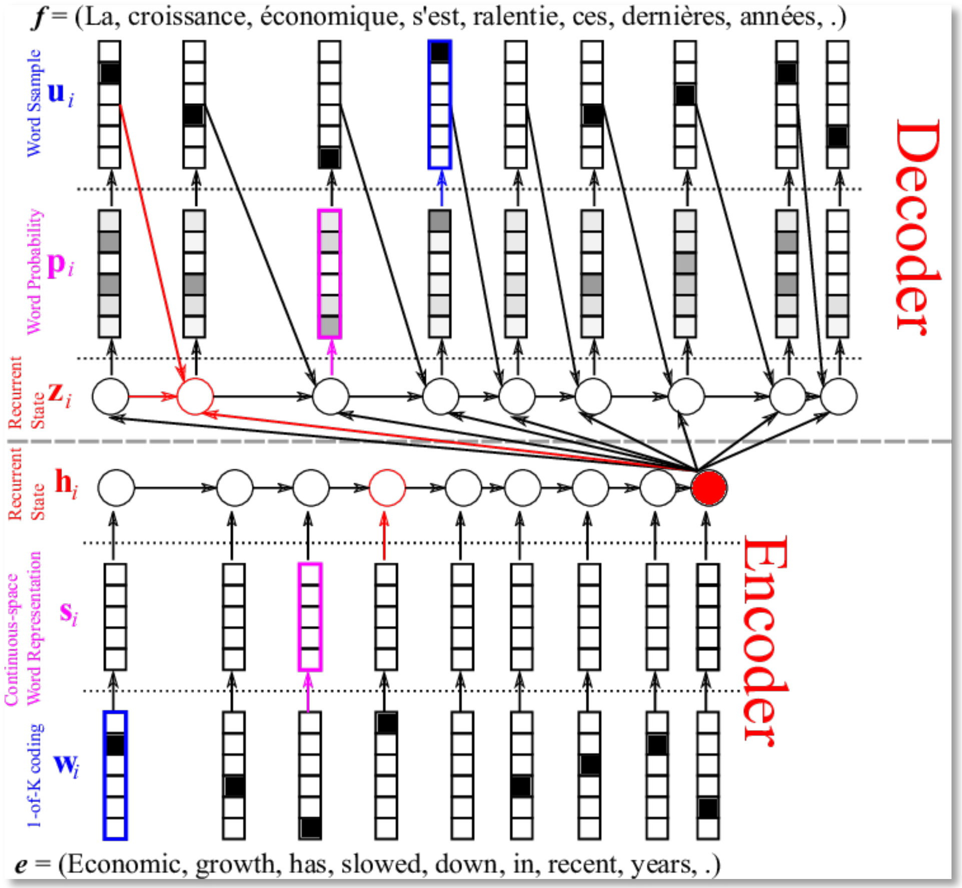
2 http://osama-oransa.blogspot.fr/2012/01/

Neural Machine Translation (NMT)

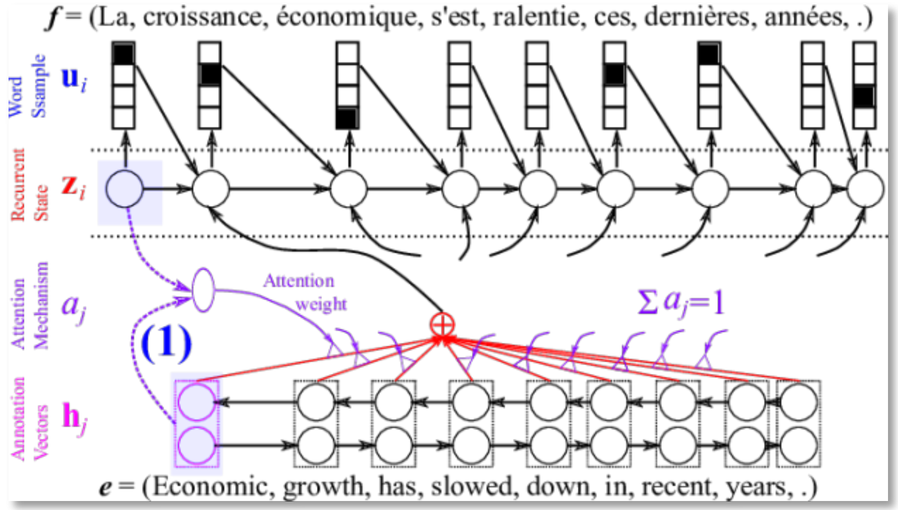
- After the recent progresses in deep learning
  - I. Sutskever & al. *Sequence to Sequence Learning with NN* (2014)
  - General end-to-end approach to sequence learning with Recurrent Neural Networks (RNNs)
  - Map input sequence to a fixed vector, decode target sequence from it
- Models later extended with attention mechanism
  - D. Bahdanau & al. *Neural Machine Translation by Jointly Learning to Align and Translate* (2014)
  - (Soft-)search parts of source relevant to predict target word

Neural Machine Translation (NMT)

- Overview[1]



- Key component: attention[1]



- Attention Mechanism takes into consideration what has been translated and one of the source words

Credit:

1 https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-with-gpus/

SMT vs. NMT

	SMT	NMT
Core element	Words	Vectors
Knowledge	Phrase table	Learned weights
Training	Slow Complex pipeline	Slower More elegant pipeline
Model size	Large	Smaller
Interpretability	Medium	Very low Opaque translation process
Introducing ling. knowledge	Doable	Doable (yet to be done!)
Open source toolkit	Yes (Moses)	Yes (many!)
Industrial deployment	Yes	Yes (now at google, systran, wipo)

… but let’s talk about performance/quality …

L. Bentivogli & al (2016) *Neural versus Phrase-Based Machine Translation Quality: a Case Study*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 257–267, Austin, Texas, November 1-5, 2016.

PAPER #1

Context

Observation

- During IWSLT 2015 shared task, NMT outperformed SMT systems on English-German pair
  - Translation of TED talks transcripts

Goal

- Analyze systems from IWSLT 2015 MT English-German task
  - A particularly challenging pair (morphology, word order)
  - 3 PBMT systems, 1 NMT system
  - Availability of post-editions of system outputs (done by professional translators)

Questions

- Strengths of NMT and weaknesses of PBMT?
- What are the linguistic phenomena that NMT handle with greater success?

Evaluation Data

- 4 systems
  - 4 sets of translation hypothesis
- Test set
  - 600 sentences (≈10K words)

System	Approach	Data
PBSY (Huck and Birch, 2015)	Combination: Phrase+Syntax-based GHKM string-to-tree; hierarchical + sparse lexicalized reordering models	175M/ 3.1B
HPB (Jehl et al., 2015)	Hierarchical Phrase-based source pre-ordering (dependency tree -based); re-scoring with neural LM	166M/ 854M
SPB (Ha et al., 2015)	Standard Phrase-based source pre-ordering (POS- and tree-based); re-scoring with neural LMs	117M/ 2.4B
NMT (Luong & Manning, 2015)	Recurrent neural network (LSTM) attention-based; source reversing; rare words handling	120M/ –

Table 1: MT systems' overview. Data column: size of parallel/monolingual training data for each system in terms of English and German tokens.

- Post-edited translations
  - minimal edits required to transform hypothesis into a fluent sentence with the same meaning as the source sentence (TER)

Translation Edit Rate (NMT is better)

- HTER (hypos/post-edits)
- mTER (hypos/closest post-edits)

system	BLEU	HTER	mTER
PBSY	25.3	28.0	21.8
HPB	24.6	29.9	23.4
SPB	25.8	29.0	22.7
NMT	31.1*	21.1*	16.2*

Table 2: Overall results on the HE Set: BLEU, computed against the original reference translation, and TER, computed with respect to the targeted post-edit (HTER) and multiple post-edits (mTER).

- * NMT is better than the score of its best competitor at statistical significance level 0.01.

Translation Quality by Sentence Length

- Results:

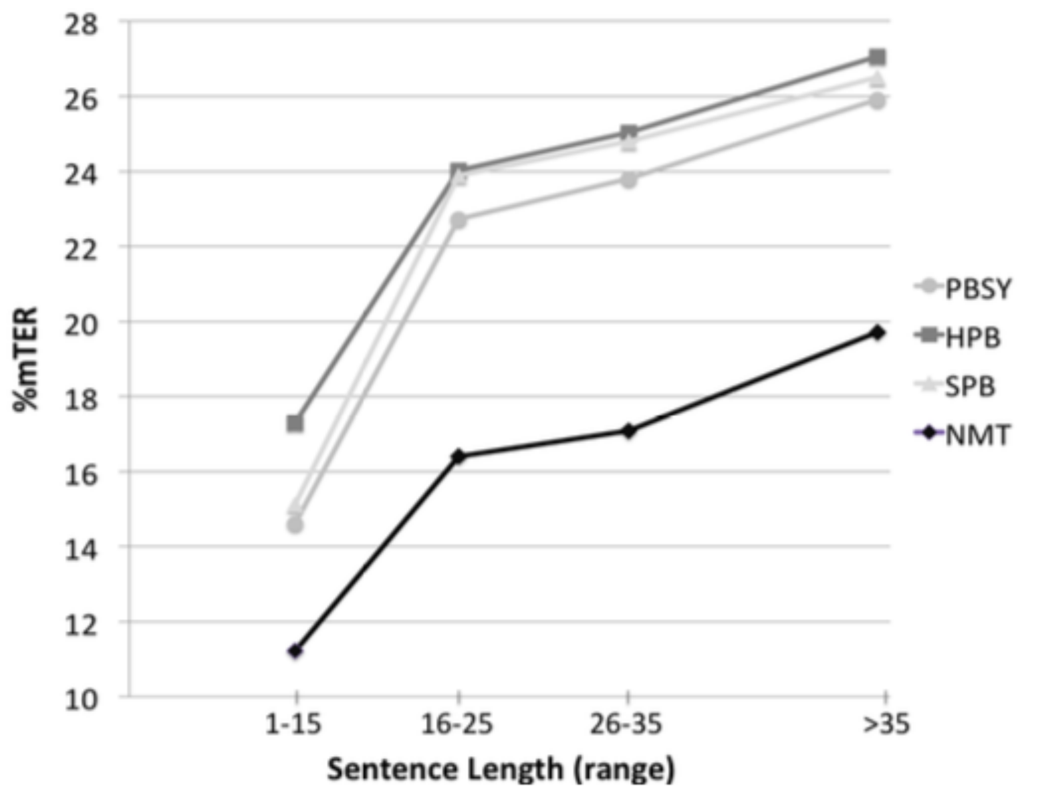


Figure 1: mTER scores on bins of sentences of different length. Points represent the average mTER of the MT outputs for the sentences in each given bin.

- **Observation**: more degradation with NMT for sentences over 35 words

Translation Quality by Talk

- Results:
- Features
  1. Length of the talk
  2. Agv. sentence length
  3. Type-Token Ratio*
     - i.e. lexical diversity
- **Observation**:
  - No correlation for features 1 & 2
  - Moderate correlation for feature 3: NMT is able to cope with lexical diversity better

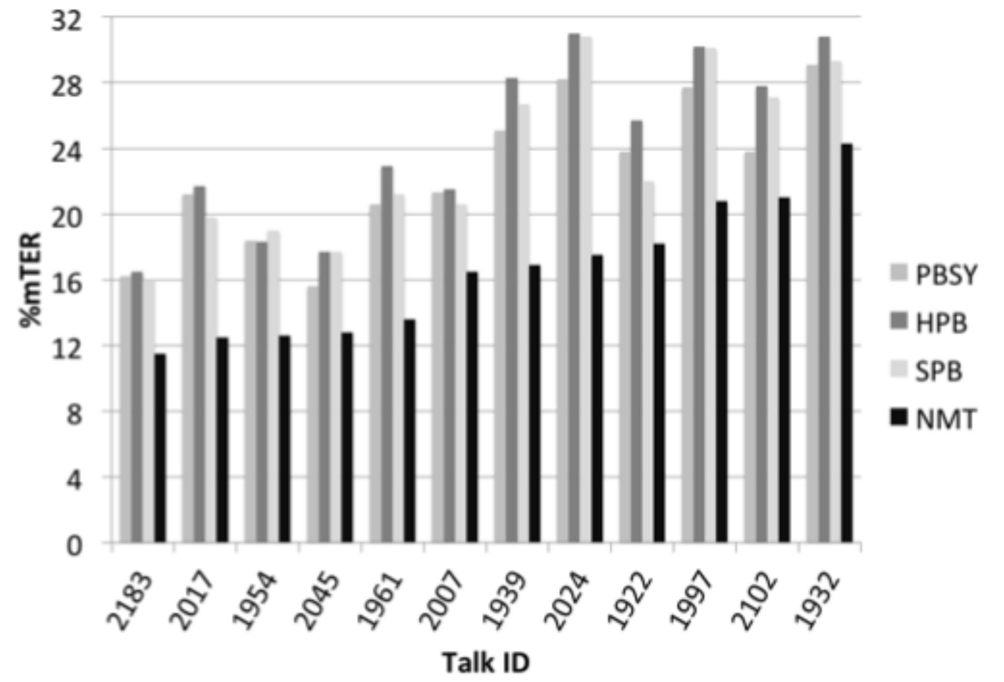


Figure 2: mTER scores per talk, sorted in ascending order of NMT scores.

* TTR of a text is calculated dividing the number of word types (vocabulary) by the total number of word tokens (occurrences)

Analysis of Translation Errors

- Three error categories:
  - *(i)* morphology errors
  - *(ii)* lexical errors
  - *(iii)* word order errors

Morphology Errors

- Results:

system	HTERnoShft		
	word	lemma	%Δ
PBSY	27.1	22.5	-16.9
HPB	28.7	23.5	-18.4
SPB	28.3	23.2	-18.0
NMT	21.7*	18.7*	-13.7

Table 3: HTER ignoring shift operations computed on words and corresponding lemmas, and their % difference.

- Computation:
  - HTER on surface forms vs HTER on lemmas: additional matches on lemmas = error on morphology
  - HTER computed without punctuation and shift position-independent ER
- **Observation**:
  - NMT generates translations which are morphologically more correct than the other systems
  - NMT makes at least 19% less morphology errors than any other PBMT system

Lexical Errors

- Computation:
  - HTER at the lemma level fits the purpose

system	HTERnoShft word	lemma	%Δ
PBSY	27.1	22.5	-16.9
HPB	28.7	23.5	-18.4
SPB	28.3	23.2	-18.0
NMT	21.7*	18.7*	-13.7

- **Observation**:
  - NMT outperforms the other systems
  - More precisely, the NMT score (18.7) is better than the second best (PBSY, 22.5) by 3.8% absolute points. This corresponds to a relative gain of about 17%, meaning that NMT makes at least 17% less lexical errors than any PBMT system
  - Similarly to what observed for morphology errors, this can be considered a remarkable improvement over the state of the art

Word Order

- Computation:
  - HTER shifts (# of words produced, # of shifts, % of shifts)
  - Kendall Reordering Score – similarity between the source-reference reorderings and the source-MT output reorderings based on words alignments

system	#words	#shifts	%shifts	KRS
PBSY	11,517	354	3.1	84.6
HPB	11,417	415	3.6	84.3
SPB	11,420	398	3.5	84.5
NMT	11,284	173	1.5*	88.3*

Table 4: Word reordering evaluation in terms of shift opera-

- Results:
- **Observation**:
  - *shift* errors in NMT translations are definitely less than in the other systems; error reduction with respect to second best (PBSY) ≈ 50% (173 *vs.* 354)
  - KRS results: the reorderings performed by NMT are much more accurate than those performed by any PBMT system

Word Order (some examples)

Auxiliary-main verb construction [aux:V]:			
	SRC	in this experiment , individuals **were shown** hundreds of hours of YouTube videos	
	HPB	in diesem Experiment , Individuen **gezeigt wurden** Hunderte von Stunden YouTube-Videos	✗
(a)	PE	in diesem Experiment **wurden** Individuen Hunderte von Stunden Youtube-Videos **gezeigt**	
	NMT	in diesem Experiment **wurden** Individuen hunderte Stunden YouTube Videos **gezeigt**	✓
	PE	in diesem Experiment **wurden** Individuen hunderte Stunden YouTube Videos **gezeigt**	
Verb in subordinate (adjunct) clause [neb:V]:			
	SRC	... when coaches and managers and owners **look** at this information streaming ...	
	PBSY	... wenn Trainer und Manager und Eigentümer **betrachten** diese Information Streaming ...	✗
(b)	PE	... wenn Trainer und Manager und Eigentümer dieses Informations-Streaming **betrachten** ...	
	NMT	... wenn Trainer und Manager und Besitzer sich diese Informationen **anschauen** ...	✓
	PE	... wenn Trainer und Manager und Besitzer sich diese Informationen **anschauen** ...	
Prepositional phrase [pp:PREP det:ART pn:N] acting as temporal adjunct:			
	SRC	so like many of us , I 've lived in a few closets **in my life**	
	SPB	so wie viele von uns , ich habe in ein paar Schränke **in meinem Leben** gelebt	✗
(c)	PE	so habe ich wie viele von uns **während meines Lebens** in einigen Verstecken gelebt	
	NMT	wie viele von uns habe ich in ein paar Schränke **in meinem Leben** gelebt	✗
	PE	wie viele von uns habe ich **in meinem Leben** in ein paar Schränken gelebt	
Negation particle [adv:PTKNEG]:			
	SRC	but I eventually came to the conclusion that that just did **not** work for systematic reasons	
	HPB	aber ich kam schlielich zu dem Schluss , dass nur aus systematischen Gründen **nicht** funktionieren	✓
(d)	PE	aber ich kam schlielich zu dem Schluss , dass es einfach aus systematischen Gründen **nicht** funktioniert	
	NMT	aber letztendlich kam ich zu dem Schluss , dass das einfach **nicht** aus systematischen Gründen funktionierte	✗
	PE	ich musste aber einsehen , dass das aus systematischen Gründen **nicht** funktioniert	

Table 6: MT output and post-edit examples showing common types of reordering errors.

Take Away Message from Paper #1

- NMT clearly outperforms SMT in term of BLEU and HTER scores
  - Even for long sentences (but NMT degrades more markedly than SMT for sent. > 35 words)
- NMT better cope with lexical diversity (moderate trend)
- NMT makes less morphology and lexical errors than SMT (moderate trend)
- Better ability to place German words (especially verbs) in the right position even when it requires considerable reordering
- NMT still struggles on more subtle translation decisions

P. Koehn & R. Knowles (2017) *Six Challenges for Neural Machine Translation*. Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver, Canada, August 4, 2017.

PAPER #2

Context

- NMT has now been deployed by Google, Systran, WIPO, etc.
  - But there have also been reports of poor performance under low-resource conditions (see DARPA LORELEI program)
- Paper examines 6 challenges to NMT based on empirical results comparing NMT (Nematus) and SMT (Moses)
  - Here we will cover 4
- Language pairs considered: English-Spanish and German-English
- Datasets from shared translation task WMT - OPUS corpus used (multi-domain)
- A 7th challenge (interpretability) is mentioned but not examined

Experimental Setup

- Language pairs
  - English–Spanish
  - German–English
- MT systems
  - SOTA NMT
    - Nematus toolkit
  - SOTA SMT
    - Moses toolkit
- Data sets
  - WMT-17: news stories
    - broad range of topic, formal language, relatively long sentences (about 30 words on average), and high standards for grammar, orthography, and style
  - Domain experiments: OPUS corpus (table 1)

Corpus	Words	Sentences	W/S
Law (Acquis)	18,128,173	715,372	25.3
Medical (EMEA)	14,301,472	1,104,752	12.9
IT	3,041,677	337,817	9.0
Koran (Tanzil)	9,848,539	480,421	20.5
Subtitles	114,371,754	13,873,398	8.2

Table 1: Corpora used to train domain-specific systems, taken from the OPUS repository. IT corpora are GNOME, KDE, PHP, Ubuntu, and OpenOffice.

Challenge 1: Domain Mismatch

- Setup
  - German–English
  - SMT
    - 5 systems trained on the opus domains + 1 system on all the training data
  - NMT
    - 5 systems trained on the opus domains + 1 system on all the training data

Challenge 1: Domain Mismatch

Results:

NMT | SMT

System ↓	Law (NMT)	Law (SMT)	Medical (NMT)	Medical (SMT)	IT (NMT)	IT (SMT)	Koran (NMT)	Koran (SMT)	Subtitles (NMT)	Subtitles (SMT)
All Data	30.5	32.8	45.1	42.2	35.3	44.7	17.9	17.9	26.4	20.8
Law	31.1	34.4	12.1	18.2	3.5	6.9	1.3	2.2	2.8	6.0
Medical	3.9	10.2	39.4	43.5	2.0	8.5	0.6	2.0	1.4	5.8
IT	1.9	3.7	6.5	5.3	42.1	39.8	1.8	1.6	3.9	4.7
Koran	0.4	1.8	0.0	2.1	0.0	2.3	15.9	18.8	1.0	5.5
Subtitles	7.0	9.9	9.3	17.8	9.2	13.6	9.0	8.4	25.9	22.1

Figure 1: Quality of systems (BLEU), when trained on one domain (rows) and tested on another domain (columns). Comparably, NMT systems (left bars) show more degraded performance out of domain.

Challenge 1: Domain Mismatch

- **Observation**:
  - In-domain NMT and SMT systems are similar (NMT is better for IT and Subtitles, SMT is better for Law, Medical, and Koran)
  - Out-of-domain performance for the NMT systems is worse in almost all cases, sometimes dramatically so
    - For instance the Medical system leads to a BLEU score of 3.9 (NMT) vs. 10.2 (SMT) on the Law test set

Challenge 1: Domain Mismatch

- Example:
- Comments
  - Careful look at NMT translation!
  - Unknown words for SMT!

Source	Schaue um dich herum.
Ref.	Look around you.
All	NMT: Look around you. SMT: Look around you.
Law	NMT: Sughum gravecorn. SMT: In order to implement dich Schaue .
Medical	NMT: EMEA / MB / 049 / 01-EN-Final Work progamme for 2002 SMT: Schaue by dich around .
IT	NMT: Switches to paused. SMT: To Schaue by itself . \t \t
Koran	NMT: Take heed of your own souls. SMT: And you see.
Subtitles	NMT: Look around you. SMT: Look around you .

Figure 2: Examples for the translation of a sentence from the Subtitles corpus, when translated with systems trained on different corpora. Performance out-of-domain is dramatically worse for NMT.

Challenge 2: Amount of Training Data

- Setup
  - English–Spanish
  - Total 385.7 million English words paired with Spanish
  - Training sets
    - 1/1024, 1/512, …, 1/2, all
- **Observation**:
  - NMT exhibits a much steeper learning curve, starting with abysmal results (1.6 vs. 16.4), outperforming SMT 25.7 vs. 24.7 with (24.1M words), and even beating the SMTsystem with a big language model with the full data set (31.1 for NMT, 28.4 for SMT, 30.4 for SMT+BigLM)

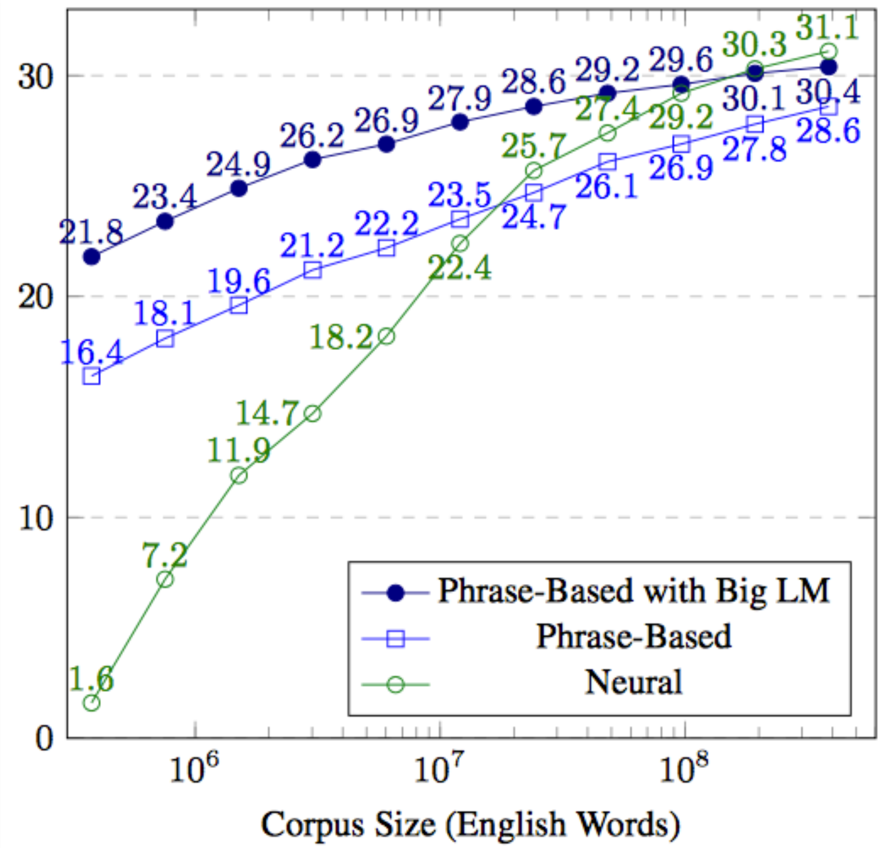


Figure 3: BLEU scores for English-Spanish systems trained on 0.4 million to 385.7 million words of parallel data. Quality for NMT starts much lower, outperforms SMT at about 15 million words, and even beats a SMT system with a big 2 billion word in-domain language model under high-resource conditions.

Challenge 3: Rare Words

- Setup
  - German–English
- **Observation**:
  - Very infrequent words
    - NMT systems actually outperform SMT systems on translation of very infrequent words
    - However, both NMT and SMT systems do continue to have difficulty translating some infrequent words, particularly those belonging to highly-inflected categories

Challenge 3: Unknown Words

- **Observation**:
  - Unknown words (not present in the training corpus)
    - The SMT system translates these correctly 53.2% of the time, while the NMT system translates them correctly 60.1% of the time
- Example:

Src.	(1) ... **choreographiertes** Gesamtkunstwerk ... (2) ... die Polizei ihn **einkesselte**.
BPE	(1) **chore@@ ograph@@ iertes** (2) **ein@@ kes@@ sel@@ te**
NMT	(1) ... **choreographed** overall artwork ... (2) ... police **stabbed** him.
SMT	(1) ... **choreographiertes** total work of art ... (2) ... police **einkesselte** him.
Ref.	(1) ... **choreographed** complete work of art ... (2) ... police **closed in on** him.

Figure 6: Examples of words that were unobserved in the training corpus, their byte-pair encodings, and their translations.

Challenge 4: Long Sentences

- Setup
  - Large English–Spanish system
  - Translation of news
  - Buckets based on source sentence length
    - 1-9, 10-19, … subwords
  - BLEU for each bucket

Challenge 4: Long Sentences

- Results:
- **Observation**:
  - Overall NMT is better than SMT
    - but the SMT system outperforms NMT on sentences of length 60 and higher.
  - Quality for the two systems is relatively close, except for the very long sentences (80 and more tokens)
    - Quality of the NMT system is dramatically lower for these since it produces too short translations (length ratio 0.859, opposed to 1.024)

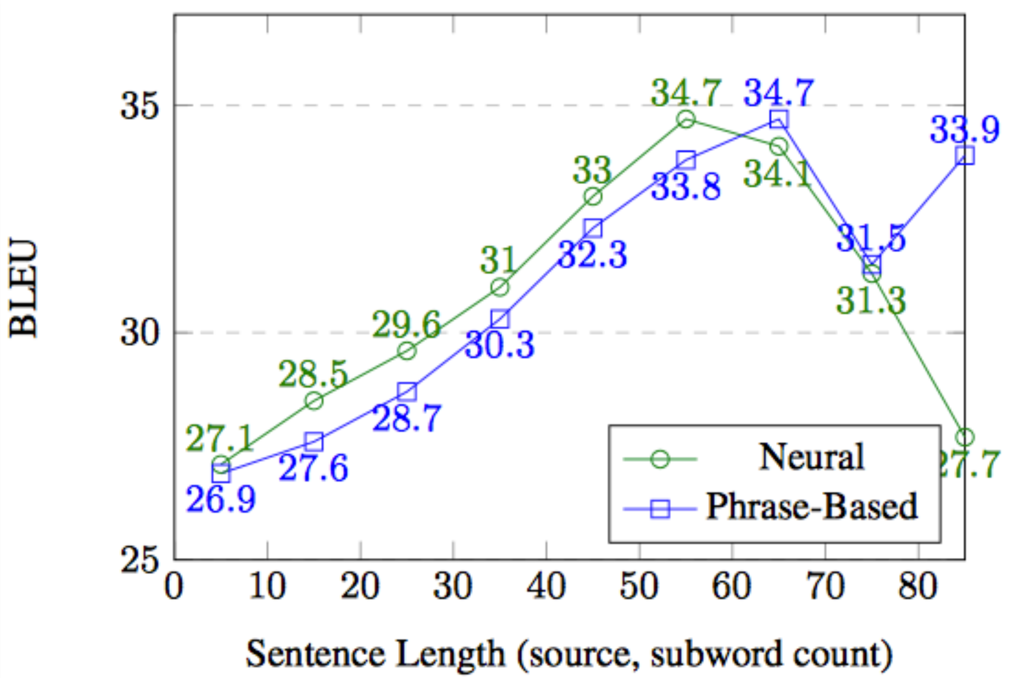


Figure 7: Quality of translations based on sentence length. SMT outperforms NMT for sentences longer than 60 subword tokens. For very long sentences (80+) quality is much worse due to too short output.

Take Away Message from Paper #2

- Out-of-domain performance of NMT is worse in almost all cases (sometimes, quite fluent outputs are totally unrelated to the input)
- NMT and SMT have very different learning curves
  - SMT is more robust in low resource conditions (< 5M words)
  - However, NMT outperforms SMT on translation of very infrequent words (use of subword units probably helps)
- While NMT trained on the full corpora is better than SMT, its quality dramatically drops for very long sentences (> 80 tokens)
- Attention model sometimes produces weird (and difficult to interpret) word alignments
- Difficult to handle large beam sizes during NMT decoding (quality drops with large search spaces)

P. Isabelle & al. (2017) *A Challenge Set Approach to Evaluating Machine Translation*. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2486–2496 Copenhagen, Denmark, September 7–11, 2017.

PAPER #3

Context

- **Observation**: Opacity of NMT systems: difficult to understand which phenomena are ill-handled by systems and why
- **Proposal**: Manual evaluation of MT on a carefully designed English dataset with difficult examples (108 sentences)
  - Each sentence in the dataset focuses on a particular linguistic phenomenon
  - Each sentence is chosen so that its closest French equivalent will be structurally divergent from the source in some crucial way
    - Morpho-syntactic divergences
    - Lexico-syntactic divergences
    - Syntactic divergences
- **Setup**: In-house (NRC) English-French SMT and NMT systems, trained on the exact same dataset, are compared
- **Distribution**: Dataset and analyses given to the community (very interesting and complete Appendix is provided)

Experimental Setup

- A set of carefully handcrafted set of 108 English sentence with their French translation
  - Language pair English–French
- Manual evaluation through yes/no questions
  - 3 bilingual native speakers rate each translated sentence

Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

Figure 1: Example challenge set question.

Experimental Setup

- MT Systems
  - SOTA MT systems trained with WMT-14 data
  - In-house PBMT
    - **PBMT-1** on the train data only (same as NMT)
      - i.e. language model from train only
    - PBMT-2 (bigger LM)
      - i.e. language model from train and mono data
  - In-house NMT (with Nematus)
    - **NMT** on train data only
  - Google's NMT
    - **GNMT**

corpus	lines	en words	fr words
train	12.1M	304M	348M
mono	15.9M	—	406M
dev	6003	138k	155k
test	3003	71k	81k

Table 1: Corpus statistics. The WMT12/13 eval sets are used for dev, and the WMT14 eval set is used for test.

Challenge Set: Divergences

- Morpho-syntactic
  - e.g. Context for subjunctive trigger
    - E: He demanded that you leave immediately.
    - F: Il a exigé que *vous partiez* immédiatement.
- Lexico-syntactic
  - e.g. Argument switching
    - E: John misses Mary
    - F: Mary *manque* à John.
  - e.g. “crossing movement” verbs
    - E: Terry swam across the river.
    - F: Terry *a traversé* la rivière *à la nage*.
      - Terry crossed the river by swimming

Challenge Set: Divergences

- Syntactic
  - e.g. position of French pronouns
    - E: He gave Mary a book. → F: Il a donné un livre à Mary.
    - E: He gave$_i$ it$_j$ to her$_k$. → F: Il *le*$_j$ *lui*$_k$ a donné$_i$.
  - e.g. stranded prepositions (WH-movement, English: preposition fronting the pronominalized object, French: preposition fronted alongside its object)
    - E: The girl whom$_i$ he was dancing with$_j$ is rich.
    - F: La fille *avec*$_j$ *qui*$_i$ il dansait est riche.
  - e.g. middle voice (English passive is agentless, not French)
    - E: Caviar is eaten with bread.
    - F: Le caviar *se mange* avec du pain.

Quantitative Comparison

- Results:

Divergence type	PBMT-1	PBMT-2	NMT	Google NMT	Agreement
Morpho-syntactic	16%	16%	72%	65%	94%
Lexico-syntactic	42%	46%	52%	62%	94%
Syntactic	33%	33%	40%	75%	81%
Overall	31%	32%	53%	68%	89%
WMT BLEU	34.2	36.5	36.9	—	—

Table 2: Summary performance statistics for each system under study, including challenge set success rate grouped by linguistic category (aggregating all positive judgments and dividing by total judgments), as well as BLEU scores on the WMT 2014 test set. The final column gives the proportion of system outputs on which all three annotators agreed.

- **Observation**:
  - Poor scores for PBMT-X, Two NMT systems clear winners
  - GNMT best overall (data & architectural factors)
  - Poor correlation with BLEU
  - Excellent interannotator agreement

Qualitative Assessment of NMT

Strengths of NMT

- Overall, both neural MT systems do much better than PBMT-1 at bridging divergences. In the case of morpho-syntactic divergences, we observe a jump from 16% to 72% in the case of our two local systems. This is mostly due to the NMT system's ability to deal with many of the more complex cases of subject-verb agreement.
- The NMT systems are also better at handling lexico-syntactic divergences.
- Finally, NMT systems also turn out to better handle purely syntactic divergences.

Weaknesses of NMT

- Globally, we note that even using a staggering quantity of data and a highly sophisticated NMT model, the Google system fails to reach the 70% mark on our challenge set.
- Here are some relevant observations:
  - *Incomplete generalizations*. In several cases where partial results might suggest that NMT has correctly captured some basic generalization about linguistic data, further instances reveals that this is not fully the case.
  - Then there are also phenomena that current NMT systems, even with massive amounts of data, appear to be completely missing: *common and syntactically flexible idioms*, *control verbs*, *argument switching verbs*, *crossing movement verbs*, and *middle voice*.

Fine-Grained Scores

Category	Subcategory	#	PBMT-1	NMT	Google NMT
Morpho-syntactic	Agreement across distractors	3	0%	100%	100%
	through control verbs	4	25%	25%	25%
	with coordinated target	3	0%	100%	100%
	with coordinated source	12	17%	92%	75%
	of past participles	4	25%	75%	75%
	Subjunctive mood	3	33%	33%	67%
Lexico-syntactic	Argument switch	3	0%	0%	0%
	Double-object verbs	3	33%	67%	100%
	Fail-to	3	67%	100%	67%
	Manner-of-movement verbs	4	0%	0%	0%
	Overlapping subcat frames	5	60%	100%	100%
	NP-to-VP	3	33%	67%	67%
	Factitives	3	0%	33%	67%
	Noun compounds	9	67%	67%	78%
	Common idioms	6	50%	0%	33%
	Syntactically flexible idioms	2	0%	0%	0%
Syntactic	Yes-no question syntax	3	33%	100%	100%
	Tag questions	3	0%	0%	100%
	Stranded preps	6	0%	0%	100%
	Adv-triggered inversion	3	0%	0%	33%
	Middle voice	3	0%	0%	0%
	Fronted should	3	67%	33%	33%
	Clitic pronouns	5	40%	80%	60%
	Ordinal placement	3	100%	100%	100%
	Inalienable possession	6	50%	17%	83%
	Zero REL PRO	3	0%	33%	100%

is the number of questions in each category

Examples: Morpho-Syntactic

S-V agreement, across distractors		
Is subject-verb agrement correct? (Possible interference from distractors between the subject's head and the verb).		
S1a	Source	The repeated calls from his mother **should** have alerted us.
	Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
	PBMT-1	Les appels répétés de sa mère aurait dû nous a alertés. ✗
	NMT	Les appels répétés de sa mère devraient nous avoir alertés. ✓
	Google	Les appels répétés de sa mère auraient dû nous alerter. ✓
S1b	Source	The sudden noise in the upper rooms **should** have alerted us.
	Ref	Le bruit soudain dans les chambres supérieures **aurait** dû nous alerter.
	PBMT-1	Le bruit soudain dans les chambres supérieures auraient dû nous a alertés. ✗
	NMT	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
	Google	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
S1c	Source	Their repeated failures to report the problem **should** have alerted us.
	Ref	Leurs échecs répétés à signaler le problème **auraient** dû nous alerter.
	PBMT-1	Leurs échecs répétés de signaler le problème aurait dû nous a alertés. ✗
	NMT	Leurs échecs répétés pour signaler le problème devraient nous avoir alertés. ✓
	Google	Leur échec répété à signaler le problème aurait dû nous alerter. ✓

Examples: Morpho-Syntactic

S-V agreement, through control verbs		
Does the flagged adjective agree correctly with its subject? (Subject-control versus object-control verbs).		
S2a	Source	She asked her brother not to be **arrogant**.
	Ref	Elle a demandé à son frère de ne pas se montrer **arrogant**.
	PBMT-1	Elle a demandé à son frère de ne pas être arrogant. ✓
	NMT	Elle a demandé à son frère de ne pas être arrogant. ✓
	Google	Elle a demandé à son frère de ne pas être arrogant. ✓
S2b	Source	She promised her brother not to be **arrogant**.
	Ref	Elle a promis à son frère de ne pas être **arrogante**.
	PBMT-1	Elle a promis son frère à ne pas être arrogant. ✗
	NMT	Elle a promis à son frère de ne pas être arrogant. ✗
	Google	Elle a promis à son frère de ne pas être arrogant. ✗
S2c	Source	She promised her doctor to remain **active** after retiring.
	Ref	Elle a promis à son médecin de demeurer **active** après s'être retirée.
	PBMT-1	Elle a promis son médecin pour demeurer actif après sa retraite. ✗
	NMT	Elle a promis à son médecin de rester actif après sa retraite. ✗
	Google	Elle a promis à son médecin de rester actif après sa retraite. ✗
S2d	Source	My mother promised my father to be more **prudent** on the road.
	Ref	Ma mère a promis à mon père d'être plus **prudente** sur la route.
	PBMT-1	Ma mère, mon père a promis d'être plus prudent sur la route. ✗
	NMT	Ma mère a promis à mon père d'être plus prudent sur la route. ✗
	Google	Ma mère a promis à mon père d'être plus prudent sur la route. ✗

Examples: Morpho-Syntactic

S-V agreement, coordinated targets		
Do the marked verbs/adjective agree correctly with their subject? (Agreement distribution over coordinated predicates)		
S3a	Source	The woman was very **tall** and extremely **strong**.
	Ref	La femme était très **grande** et extrêmement **forte**.
	PBMT-1	La femme était très gentil et extrêmement forte. ✗
	NMT	La femme était très haute et extrêmement forte. ✓
	Google	La femme était très grande et extrêmement forte. ✓
S3b	Source	Their politicians were more **ignorant** than **stupid**.
	Ref	Leurs politiciens étaient plus **ignorants** que **stupides**.
	PBMT-1	Les politiciens étaient plus ignorants que stupide. ✗
	NMT	Leurs politiciens étaient plus ignorants que stupides. ✓
	Google	Leurs politiciens étaient plus ignorants que stupides. ✓
S3c	Source	We **shouted** an insult and **left** abruptly.
	Ref	Nous **avons** lancé une insulte et nous **sommes** partis brusquement.
	PBMT-1	Nous avons crié une insulte et a quitté abruptement. ✗
	NMT	Nous avons crié une insulte et nous avons laissé brusquement. ✓
	Google	Nous avons crié une insulte et nous sommes partis brusquement. ✓

Examples: Morpho-Syntactic

S-V agreement, feature calculus on coordinated source		
Do the marked verbs/adjective agree correctly with their subject? (Masculine singular ET masculine singular yields masculine plural).		
S4a1	Source	The cat and the dog **should** be **watched**.
	Ref	Le chat et le chien **devraient** être **surveillés**.
	PBMT-1	Le chat et le chien doit être regardée. ✗
	NMT	Le chat et le chien doivent être regardés. ✓
	Google	Le chat et le chien doivent être surveillés. ✓
S4a2	Source	My father and my brother **will** be **happy** tomorrow.
	Ref	Mon père et mon frère **seront heureux** demain.
	PBMT-1	Mon père et mon frère sera heureux de demain. ✗
	NMT	Mon père et mon frère seront heureux demain. ✓
	Google	Mon père et mon frère seront heureux demain. ✓
S4a3	Source	My book and my pencil **could** be **stolen**.
	Ref	Mon livre et mon crayon **pourraient** être **volés**.
	PBMT-1	Mon livre et mon crayon pourrait être volé. ✗
	NMT	Mon livre et mon crayon pourraient être volés. ✓
	Google	Mon livre et mon crayon pourraient être volés. ✓

Examples: Lexico-Syntactic

Argument switch		
Are the experiencer and the object of the "missing" situation correctly preserved in the French translation? (Argument switch).		
S7a	Source	**Mary** sorely misses **Jim**.
	Ref	**Jim** manque cruellement **à Mary**.
	PBMT-1	Marie manque cruellement de Jim. ✗
	NMT	Mary a lamentablement manqué de Jim. ✗
	Google	Mary manque cruellement à Jim. ✗
S7b	Source	**My sister** is really missing **New York.**
	Ref	**New York** manque beaucoup **à ma sœur**.
	PBMT-1	Ma sœur est vraiment absent de New York. ✗
	NMT	Ma sœur est vraiment manquante à New York. ✗
	Google	Ma sœur manque vraiment New York. ✗
S7c	Source	What **he** misses most is **his dog**.
	Ref	Ce qui **lui** manque le plus, c'est **son chien**.
	PBMT-1	Ce qu'il manque le plus, c'est son chien. ✗
	NMT	Ce qu'il manque le plus, c'est son chien. ✗
	Google	Ce qu'il manque le plus, c'est son chien. ✗

Examples: Lexico-Syntactic

Manner-of-movement verbs		
Is the movement action expressed in the English source correctly rendered in French? (Manner-of-movement verbs with path argument may need to be rephrased in French).		
S10a	Source	John would like to **swim across** the river.
	Ref	John aimerait **traverser** la rivière **à la nage**.
	PBMT-1	John aimerait nager dans la rivière. ✗
	NMT	John aimerait nager à travers la rivière. ✗
	Google	John aimerait nager à travers la rivière. ✗
S10b	Source	They **ran into** the room.
	Ref	Ils **sont entrés** dans la chambre **à la course**.
	PBMT-1	Ils ont couru dans la chambre. ✗
	NMT	Ils ont couru dans la pièce. ✗
	Google	Ils coururent dans la pièce. ✗
S10c	Source	The man **ran out of** the park.
	Ref	L'homme **est sorti du** parc **en courant**.
	PBMT-1	L'homme a manqué du parc. ✗
	NMT	L'homme s'enfuit du parc. ✗
	Google	L'homme sortit du parc. ✗
Hard example featuring spontaneous noun-to-verb derivation ("nonce verb").		
S10d	Source	John **guitared his way** to San Francisco.
	Ref	John **s'est rendu** jusqu'à San Francisco **en jouant de la guitare**.
	PBMT-1	John guitared son chemin à San Francisco. ✗
	NMT	John guitared sa route à San Francisco. ✗
	Google	John a guité son chemin à San Francisco. ✗

Examples: Lexico-Syntactic

Overlapping subcat frames		
Is the French verb for "know" correctly chosen? (Choice between "savoir"/"connaître" depends on syntactic nature of its object)		
S11a	Source	Paul **knows** that this is a fact.
	Ref	Paul **sait que** c'est un fait.
	PBMT-1	Paul sait que c'est un fait. ✓
	NMT	Paul sait que c'est un fait. ✓
	Google	Paul sait que c'est un fait. ✓
S11b	Source	Paul **knows** this story.
	Ref	Paul **connaît** cette histoire.
	PBMT-1	Paul connaît cette histoire. ✓
	NMT	Paul connaît cette histoire. ✓
	Google	Paul connaît cette histoire. ✓
S11c	Source	Paul **knows** this story is hard to believe.
	Ref	Paul **sait que** cette histoire est difficile à croire.
	PBMT-1	Paul connaît cette histoire est difficile à croire. ✗
	NMT	Paul sait que cette histoire est difficile à croire. ✓
	Google	Paul sait que cette histoire est difficile à croire. ✓
S11d	Source	He **knows** my sister will not take it.
	Ref	Il **sait que** ma soeur ne le prendra pas.
	PBMT-1	Il sait que ma soeur ne prendra pas. ✓
	NMT	Il sait que ma soeur ne le prendra pas. ✓
	Google	Il sait que ma soeur ne le prendra pas. ✓
S11e	Source	My sister **knows** your son is reliable.
	Ref	Ma sœur **sait que** votre fils est fiable.
	PBMT-1	Ma soeur connaît votre fils est fiable. ✗
	NMT	Ma sœur sait que votre fils est fiable. ✓
	Google	Ma sœur sait que votre fils est fiable. ✓

Examples: Lexico-Syntactic

Common idioms		
Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?		
S15a	Source	Stop **beating around the bush**.
	Ref	Cessez de **tourner autour du pot**.
	PBMT-1	Cesser de battre la campagne. ✗
	NMT	Arrêtez de battre autour de la brousse. ✗
	Google	Arrêter de tourner autour du pot. ✓
S15b	Source	You are **putting the cart before the horse**.
	Ref	Vous **mettez la charrue devant les bœufs**.
	PBMT-1	Vous pouvez mettre la charrue avant les bœufs. ✓
	NMT	Vous mettez la charrue avant le cheval. ✗
	Google	Vous mettez le chariot devant le cheval. ✗
S15c	Source	His comment proved to be **the straw that broke the camel's back**.
	Ref	Son commentaire s'est avéré être **la goutte d'eau qui a fait déborder le vase**.
	PBMT-1	Son commentaire s'est révélé être la goutte d'eau qui fait déborder le vase. ✓
	NMT	Son commentaire s'est avéré être la paille qui a brisé le dos du chameau. ✗
	Google	Son commentaire s'est avéré être la paille qui a cassé le dos du chameau. ✗
S15d	Source	His argument really **hit the nail on the head**.
	Ref	Son argument a vraiment **fait mouche**.
	PBMT-1	Son argument a vraiment mis le doigt dessus. ✓
	NMT	Son argument a vraiment frappé le clou sur la tête. ✗
	Google	Son argument a vraiment frappé le clou sur la tête. ✗
S15e	Source	It's **no use crying over spilt milk**.
	Ref	**Ce qui est fait est fait**.
	PBMT-1	Ce n'est pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait haché. ✗
	Google	Ce qui est fait est fait. ✓
S15f	Source	It is **no use crying over spilt milk**.
	Ref	**Ce qui est fait est fait**.
	PBMT-1	Il ne suffit pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait écrémé. ✗
	Google	Il est inutile de pleurer sur le lait répandu. ✗

Examples: Lexico-Syntactic

Syntactically flexible idioms		
Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?		
S16a	Source	The cart has been put before the horse.
	Ref	La **charrue a été mise devant les bœufs**.
	PBMT-1	On met la charrue devant le cheval. ✗
	NMT	Le chariot a été mis avant le cheval. ✗
	Google	Le chariot a été mis devant le cheval. ✗
S16b	Source	With this argument, **the nail has been hit on the head**.
	Ref	Avec cet argument, **la cause est entendue**.
	PBMT-1	Avec cette argument, l'ongle a été frappée à la tête. ✗
	NMT	Avec cet argument, l'ongle a été touché à la tête. ✗
	Google	Avec cet argument, le clou a été frappé sur la tête. ✗

Examples: Syntactic

Yes-no question syntax		
Is the English question correctly rendered as a French question?		
S17a	Source	**Have the kids** ever watched that movie?
	Ref	**Les enfants ont-ils** déjà vu ce film?
	PBMT-1	Les enfants jamais regardé ce film? ✗
	NMT	Les enfants ont-ils déjà regardé ce film? ✓
	Google	Les enfants ont-ils déjà regardé ce film? ✓
S17b	Source	**Hasn't your boss denied you** a promotion?
	Ref	**Votre patron ne vous a-t-il pas refusé** une promotion?
	PBMT-1	N'a pas nié votre patron vous un promotion? ✗
	NMT	Est-ce que votre patron vous a refusé une promotion? ✓
	Google	Votre patron ne vous a-t-il pas refusé une promotion? ✓
S17c	Source	**Shouldn't I attend** this meeting?
	Ref	**Ne devrais-je pas assister** à cette réunion?
	PBMT-1	Ne devrais-je pas assister à cette réunion? ✓
	NMT	Est-ce que je ne devrais pas assister à cette réunion? ✓
	Google	Ne devrais-je pas assister à cette réunion? ✓

Examples: Syntactic

Tag questions		
Is the English "tag question" element correctly rendered in the translation?		
S18a	Source	Mary looked really happy tonight, **didn't she**?
	Ref	Mary avait l'air vraiment heureuse ce soir, **n'est-ce pas**?
	PBMT-1	Marie a regardé vraiment heureux de ce soir, n'est-ce pas elle? ✗
	NMT	Mary s'est montrée vraiment heureuse ce soir, ne l'a pas fait? ✗
	Google	Mary avait l'air vraiment heureuse ce soir, n'est-ce pas? ✓
S18b	Source	We should not do that again, **should we**?
	Ref	Nous ne devrions pas refaire cela, **n'est-ce pas**?
	PBMT-1	Nous ne devrions pas faire qu'une fois encore, faut-il? ✗
	NMT	Nous ne devrions pas le faire encore, si nous? ✗
	Google	Nous ne devrions pas recommencer, n'est-ce pas? ✓
S18c	Source	She was perfect tonight, **was she not**?
	Ref	Elle était parfaite ce soir, **n'est-ce pas**?
	PBMT-1	Elle était parfait ce soir, elle n'était pas? ✗
	NMT	Elle était parfaite ce soir, n'était-elle pas? ✗
	Google	Elle était parfaite ce soir, n'est-ce pas? ✓

Examples: Syntactic

WH-MVT and stranded preps

Is the dangling preposition of the English sentence correctly placed in the French translation?

S19a	Source	The guy **that** she is going out **with** is handsome.
	Ref	Le type **avec qui** elle sort est beau.
	PBMT-1	Le mec qu'elle va sortir avec est beau. ✗
	NMT	Le mec qu'elle sort avec est beau. ✗
	Google	Le mec avec qui elle sort est beau. ✓
S19b	Source	**Whom** is she going out **with** these days?
	Ref	**Avec qui** sort-elle ces jours-ci?
	PBMT-1	Qu'est-ce qu'elle allait sortir avec ces jours? ✗
	NMT	À qui s'adresse ces jours-ci? ✗
	Google	Avec qui sort-elle de nos jours? ✓
S19c	Source	The girl **that** he has been talking **about** is smart.
	Ref	La fille **dont** il a parlé est brillante.
	PBMT-1	La jeune fille qu'il a parlé est intelligent. ✗
	NMT	La fille qu'il a parlé est intelligente. ✗
	Google	La fille dont il a parlé est intelligente. ✓
S19d	Source	**Who** was he talking **to** when you left?
	Ref	**À qui** parlait-il au moment où tu es parti?
	PBMT-1	Qui est lui parler quand vous avez quitté? ✗
	NMT	Qui a-t-il parlé à quand vous avez quitté? ✗
	Google	Avec qui il parlait quand vous êtes parti? ✓
S19e	Source	The city **that** he is arriving **from** is dangerous.
	Ref	La ville **d'où** il arrive est dangereuse.
	PBMT-1	La ville qu'il est arrivé de est dangereuse. ✗
	NMT	La ville qu'il est en train d'arriver est dangereuse. ✗
	Google	La ville d'où il vient est dangereuse. ✓
S19f	Source	**Where** is he arriving **from**?
	Ref	**D'où** arrive-t-il?
	PBMT-1	Où est-il arrivé? ✗
	NMT	De quoi s'agit-il? ✗
	Google	D'où vient-il? ✓

Examples: Syntactic

Middle voice		
Is the generic statement made in the English sentence correctly and naturally rendered in the French translation?		
S21a	Source	Soup **is eaten** with a large spoon.
	Ref	La soupe **se mange** avec une grande cuillère
	PBMT-1	La soupe est mangé avec une grande cuillère. ✗
	NMT	La soupe est consommée avec une grosse cuillère. ✗
	Google	La soupe est consommée avec une grande cuillère. ✗
S21b	Source	Masonry **is cut** using a diamond blade.
	Ref	La maçonnerie **se coupe** avec une lame à diamant.
	PBMT-1	La maçonnerie est coupé à l'aide d'une lame de diamant. ✗
	NMT	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
	Google	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
S21c	Source	Champagne **is drunk** in a glass called a flute.
	Ref	Le champagne **se boit** dans un verre appelé flûte.
	PBMT-1	Le champagne est ivre dans un verre appelé une flûte. ✗
	NMT	Le champagne est ivre dans un verre appelé flûte. ✗
	Google	Le Champagne est bu dans un verre appelé flûte. ✗

Examples: Syntactic

Fronted "should"		
Fronted "should" is interpreted as a conditional subordinator. It is normally translated as "si" with imperfect tense.		
S22a	Source	**Should** Paul leave, I would be sad.
	Ref	**Si** Paul **devait** s'en aller, je serais triste.
	PBMT-1	Si le congé de Paul, je serais triste. ✗
	NMT	Si Paul quitte, je serais triste. ✗
	Google	Si Paul s'en allait, je serais triste. ✓
S22b	Source	Should he become president, she would be promoted immediately.
	Ref	**S'**il devait devenir président, elle recevrait immédiatement une promotion.
	PBMT-1	S'il devait devenir président, elle serait encouragée immédiatement. ✓
	NMT	S'il devait devenir président, elle serait immédiatement promue. ✓
	Google	Devrait-il devenir président, elle serait immédiatement promue. ✗
S22c	Source	**Should** he fall, he would get up again immediately.
	Ref	**S'** il venait à tomber, il se relèverait immédiatement.
	PBMT-1	S'il devait tomber, il allait se lever immédiatement de nouveau. ✓
	NMT	S'il tombe, il serait de nouveau immédiatement. ✗
	Google	S'il tombe, il se lèvera immédiatement. ✗

Examples: Syntactic

Inalienable possession		
Is the French translation correct and natural both in: a) its use of a particular determiner on the body part noun; and b) the presence or absence of a reflexive pronoun before the verb?		
S25a	Source	He washed **his** hands.
	Ref	Il **s'est** lavé **les** mains.
	PBMT-1	Il se lavait les mains. ✓
	NMT	Il a lavé ses mains. ✗
	Google	Il se lava les mains. ✓
S25b	Source	I brushed **my** teeth.
	Ref	Je **me** suis brossé **les** dents.
	PBMT-1	J'ai brossé mes dents. ✗
	NMT	J'ai brossé mes dents. ✗
	Google	Je me suis brossé les dents. ✓
S25c	Source	You brushed **your** teeth.
	Ref	Tu **t'**es brossé **les** dents
	PBMT-1	Vous avez brossé vos dents. ✗
	NMT	vous avez brossé vos dents. ✗
	Google	Tu as brossé les dents. ✗
S25d	Source	I raised **my** hand.
	Ref	J'ai levé **la** main.
	PBMT-1	J'ai levé la main. ✓
	NMT	J'ai soulevé ma main. ✗
	Google	Je levai la main. ✓
S25e	Source	He turned **his** head.
	Ref	Il a tourné **la** tête.
	PBMT-1	Il a transformé sa tête. ✗
	NMT	Il a tourné sa tête. ✗
	Google	Il tourna la tête. ✓

Examples: Syntactic

Zero REL PRO

Is the English zero relative pronoun correctly translated as a non-zero one in the French translation?

S26a	Source	The strangers the woman saw were working.
	Ref	Les inconnus **que** la femme vit travaillaient.
	PBMT-1	Les étrangers la femme vit travaillaient. ✗
	NMT	Les inconnus de la femme ont travaillé. ✗
	Google	Les étrangers que la femme vit travaillaient. ✓
S26b	Source	The man your sister hates is evil.
	Ref	L'homme **que** votre sœur déteste est méchant.
	PBMT-1	L'homme ta soeur hait est le mal. ✗
	NMT	L'homme que ta soeur est le mal est le mal. ✓
	Google	L'homme que votre sœur hait est méchant. ✓
S26c	Source	The girl my friend was talking about is gone.
	Ref	La fille **dont** mon ami parlait est partie.
	PBMT-1	La jeune fille mon ami a parlé a disparu. ✗
	NMT	La petite fille de mon ami était révolue. ✗
	Google	La fille dont mon ami parlait est partie. ✓

Take Away Message from Paper #3

- SMT systems do poorly on the challenge set (NMT is better) while BLEU scores of both systems are similar for WMT shared task
- NMT better than SMT at bridging divergences
- Gap between in-house (NRC) and commercial (Google) NMT results suggests that, given enough data, NMT systems can successfully tackle difficult challenges
- NMT has still serious shortcomings (incomplete list)
  - Noun compounds (N1 N2 => N2 prep N1)
  - Common and syntactically flexible idioms
  - Argument switching verbs (N1 misses N2 => N2 manque à N1)
  - Crossing movement verbs (swim across X => traverser X à la nage)

For more

http://www.lemonde.fr/sciences/video/2017/06/30/les-defis-de-la-traduction-automatique_5153681_1650684.html