A Stratified sampling of the Hamming distance

Norouzi et al. (2016) detail the steps to drawing samples from the reward distribution based on the edit distance with stratified sampling. In this section we show how we sample from the Hamming distance (a special case of the edit distance) reward and how we handle large vocabularies to generate reasonable candidates. To draw from the Hamming distance reward, we proceed as follows:

1. Sample a distance d from $\{0, \dots, T\}$.
2. Pick d positions in the sequence to be changed among $\{1, \dots, T\}$.
3. Sample substitutions from a subset $\mathcal{V}_{sub}$ of the vocabulary ($|\mathcal{V}_{sub}| = V_{sub}$).

To sample a distance, we partition the set of sequences in $\mathcal{V}_{sub}$ terms $\mathcal{V}_{sub}^T$ with respect to their distance to the ground truth y^*:

$$\begin{cases} S_d = \{y \in \mathcal{V}_{sub}^T | \, d(y, y^*) = d\}, \\ \mathcal{V}_{sub}^T = \underset{d}{\cup} S_d, \\ \forall d, d' : S_d \cap S_{d'} = \emptyset. \end{cases}$$

To each distance d in $\{0, ..., T\}$ we assign the portion of rewards covered by S_d in $\mathcal{V}_{sub}^T$. Since all elements of S_d are assigned the same reward $e^{-\frac{d}{\tau}}$; we need only to multiply it by the set's size $|S_d|$. Counting the elements of $|S_d|$ is straight-forward: we choose d elements to alter in y^* $\left(\binom{T}{d}\text{combinations}\right)$ and at each position we have $(V_{sub} - 1)$ possibilities. The sampling distribution is obtained as:

$$p(d) = r(S_d)/r(\mathcal{V}_{sub}^T) \tag{20}$$

$$= \frac{\sum\limits_{y \in S_d} r(y|y^*)}{\sum\limits_{y \in \mathcal{V}_{sub}^T} r(y|y^*)}. \tag{21}$$

Given that $\{S_d\}_d$ form a partition of $\mathcal{V}_{sub}^T$,

$$\sum_{y \in \mathcal{V}_{sub}^T} r(y|y^*) = \sum_d \sum_{y \in S_d} e^{-\frac{d}{\tau}} \tag{22}$$

$$= \sum_d \binom{T}{d} (V_{sub} - 1)^d e^{-\frac{d}{\tau}} \tag{23}$$

$$= \left((V_{sub} - 1)e^{-\frac{1}{\tau}} + 1\right)^T, \tag{24}$$

we find:

$$p(d) = \binom{T}{d} \frac{(V_{sub} - 1)^d e^{-\frac{d}{\tau}}}{((V_{sub} - 1)e^{-\frac{1}{\tau}} + 1)^T}. \tag{25}$$

B Captioning

B.1 Experimental setup

Out of vocabulary words are replaced by <UNK> token and the captions longer than 16 words are truncated. As image encoding, we average-pool the features in the last convolutional layer of ResNet-152 (He et al., 2016) pre-trained on ImageNet. The 2048-dimensional image signature is further mapped to $\mathbb{R}^{512}$ to fit the word-embedding dimension, so it can be used as the first token fed to the RNN decoder. We use a single-layer RNN with $d = 512$ LSTM units.

For optimization, we use Adam (Kingma and Ba, 2015) with a batch size of 10 images, *i.e.* 50 sentences. We follow Lu et al. (2017); Pedersoli et al. (2017) and train in two stages: the first, optimizing

the language model alone with an initial learning rate of 5e-4 annealed by a factor of 0.6 every 3 epochs starting from the 5th one. We train for up to 20 epochs with early stopping if CIDEr score on the validation set does not improve. In the second stage, we optimize the language model conjointly with $conv_4$ and $conv_5$ (the last 39 building blocks of ResNet-152) of the CNN model. The initial learning rate is of 6e-5 and diminishes by a factor of 0.8 every 4 epochs. The same early-stopping strategy is applied. For the token-level reward, we use GloVe (Pennington et al., 2014) as our word embedding, which we train on the captions in the MS-COCO training set. In preliminary experiments using the publicly available 300-dimensional GloVe vectors trained on Wikipedia 2014 + Gigaword worsens the model's results.

B.2 Restricted vocabulary sampling - supplementary results

In Table 4 we provide additional results for sequence-level smoothing, when using the BLEU-4 as reward function. We include results when computing BLEU-4 w.r.t. all reference sequences (like also done for CIDEr), and when computing w.r.t. a single randomly selected reference sentence (as is the case for Hamming). With the BLEU-4 reward, using $\mathcal{V}_{refs}$ yields best results in all but a single case. This underlines the effectiveness of sampling replacement words that are relevant to the task in sequence-level smoothing.

Captioning without attention					**Captioning with attention**				
Reward	$\mathcal{V}_{sub}$	BLEU-1	BLEU-4	CIDEr	Reward	$\mathcal{V}_{sub}$	BLEU-1	BLEU-4	CIDEr
Dirac		70.63	30.14	93.59	Dirac		73.40	33.11	101.63
Hamming	$\mathcal{V}$	71.76	31.16	96.37	Hamming	$\mathcal{V}$	73.12	32.71	101.25
Hamming	$\mathcal{V}_{batch}$	71.46	31.15	**96.53**	Hamming	$\mathcal{V}_{batch}$	73.26	**32.73**	101.90
Hamming	$\mathcal{V}_{refs}$	**71.80**	**31.63**	96.22	Hamming	$\mathcal{V}_{refs}$	**73.53**	32.59	**102.33**
BLEU-4 single	$\mathcal{V}$	71.30	31.11	96.11	BLEU-4 single	$\mathcal{V}$	72.98	32.55	101.15
BLEU-4 single	$\mathcal{V}_{batch}$	71.13	30.84	94.74	BLEU-4 single	$\mathcal{V}_{batch}$	72.98	32.48	101.05
BLEU-4 single	$\mathcal{V}_{refs}$	**71.78**	**31.63**	**97.29**	BLEU-4 single	$\mathcal{V}_{refs}$	**73.41**	**32.69**	**101.35**
BLEU-4	$\mathcal{V}$	71.56	**31.47**	96.56	BLEU-4	$\mathcal{V}$	73.39	32.89	102.60
BLEU-4	$\mathcal{V}_{batch}$	71.41	30.87	95.69	BLEU-4	$\mathcal{V}_{batch}$	73.24	32.74	102.49
BLEU-4	$\mathcal{V}_{refs}$	**72.08**	31.41	**97.21**	BLEU-4	$\mathcal{V}_{refs}$	**73.55**	**33.03**	**102.72**
CIDEr	$\mathcal{V}$	71.05	30.46	94.40	CIDEr	$\mathcal{V}$	73.08	32.51	101.84
CIDEr	$\mathcal{V}_{batch}$	71.51	31.17	95.78	CIDEr	$\mathcal{V}_{batch}$	**73.50**	**33.04**	**102.98**
CIDEr	$\mathcal{V}_{refs}$	**71.93**	**31.41**	**96.81**	CIDEr	$\mathcal{V}_{refs}$	73.42	32.91	102.23

Table 4: Captioning performance on MSCOCO when training with sequence-level loss smoothing.

B.3 Training time

We report below (Table 5) the average wall time to process a single batch (10 images *i.e.* 50 captions) when training the RNN language model with fixed CNN (without attention) on a Titan X GPU. We can clearly see that the lazy training is faster compared to the standard sequence smoothing and that the token-level smoothing does not hinder the training speed.

Loss	MLE	Tok	Seq	Seq lazy	Seq	Seq lazy	Seq	Seq lazy	Tok-Seq	Tok-Seq	Tok-Seq
Reward		Glove sim	Hamming								
$\mathcal{V}_{sub}$			$\mathcal{V}$	$\mathcal{V}$	$\mathcal{V}_{batch}$	$\mathcal{V}_{batch}$	$\mathcal{V}_{refs}$	$\mathcal{V}_{refs}$	$\mathcal{V}$	$\mathcal{V}_{batch}$	$\mathcal{V}_{refs}$
ms/batch	347	359	390	349	395	337	401	336	445	446	453

Table 5: Average training time per batch for different losses

B.4 Additional examples

Ground truth:
a man holding a sausage dog and looking at the sausage do
a man in a suit stares at a chili dog with cheese
a man is eating a hot dog while wearing a suit
a man looks at a hot dog he is eating
a man holds up a partially eaten hotdog

Generated:
Baseline: a man holding a donut in his hand
Seq: a man eating a donut in a restaurant
Tok: a man holding a sandwich in his hands
Tok-Seq: a man holding a hot dog in his hand

Ground truth:
dirt bikes with lights on riding along a street with people watching from the side walk
the motorcycle procession made their way down the crowded street
police officers are on parade as crowds watch from aside
a group of motorcycles are going down the road
a line of police motorcycles driving down the road

Generated:
Baseline: a man riding a motorcycle down a street
Seq: a group of people riding bikes down a street
Tok: a group of people riding motorcycles down a street
Tok-Seq: a group of people riding motorcycles down a street

Ground truth:
a vase with a flower sitting on top of a wooden table
a flower vase with a glow to it sitting on a table
a small, illuminated vase holds a single tulip at a cafe
a weird and strange looking vase in a flower.
a single tulip is seen in a small vase

Generated:
Baseline: a glass vase with a candle in it
Seq: a glass vase sitting on top of a table
Tok: a vase with a flower in it on a table
Tok-Seq: a glass vase with a flower in it

Ground truth:
a background of blurred shapes is fronted by bunches of green bananas of which one's been
a bunch of small bananas still on the main branch
a large bunch of unripened green plantains on display
a picture of a bunches of green bananas
a bunch of bananas are piled together

Generated:
Baseline: a bunch of green bananas hanging from a tree
Seq: a bunch of green bananas on a tree
Tok: a bunch of green bananas hanging from a tree
Tok-Seq: a bunch of green bananas sitting on a table

Ground truth:
a man standing in front of a street vendor, a woman behind the counter
a man preparing food on a street cart next to a building
a man running a street stand in front of a yellow garag
a man and a woman stand at a portable vending cart
a woman eats a burger next to a street vendor

Generated:
Baseline: a man is standing next to a bicycle
Seq: a man on a bike with a man on the back
Tok: a couple of men standing next to each other
Tok-Seq: a man standing next to a woman in front of a store

Ground truth:
a colorful walk sign in the city is on a post
an intersection cross walk at greene street featuring the 'walk' ligh
view of street signs and an illuminated cross walk sign
the crosswalk sign is placed under a street sign
traffic signal for saying walk on greene st

Generated:
Baseline: a traffic light with a street sign on it
Seq: a street sign with a street sign on it
Tok: a traffic light and street sign on a pole
Tok-Seq: a traffic light and street sign on a pole

Ground truth:
there are two men standing on the beach with surf board
the two surfers are read to take on the waves
two surfers are standing on a beach holding their surfboards
a pair of surfers pause on a populated beach
two surfers with surfboards walk upon a beach

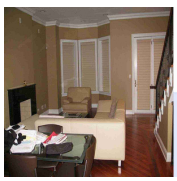
Generated:
Baseline: a couple of people standing on top of a beach
Seq: a couple of people that are on a beach
Tok: a couple of people standing on top of a beach
Tok-Seq: a couple of people that are holding surfboards

Ground truth:
a display case filled with bakes goods in front of a store
the front of a chinese store with some items on display
a chinese restaurant with a bunch of different plants near it.
a photo of an asian market with a display case
the counter of an ethnic asian cuisine bar

Generated:
Baseline: a display case filled with different types of donuts
Seq: a store filled with lots of fruit and vegetables
Tok: a display case filled with lots of donuts
Tok-Seq: a display case filled with lots of flowers

Ground truth:
the blossom of a plant hangs near large green leaves
small bananas growing on top of the banana tre
a weird looking tree filled with little miniature bananas
a large green leafy plant with a flower
a tree with tiny bananas growing on it

Generated:
Baseline: a close up of a bunch of green bananas
Seq: a tree with a bunch of bananas growing in it
Tok: a close up of a banana plant with leaves
Tok-Seq: a tree filled with lots of green bananas

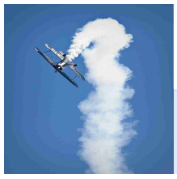


Ground truth:
a living room with a sectional couch, easy chair and glass desk covered in paper
home living room with brown walls with white trim, fireplace, and tan furnishings
a living room includes a beige sofa and a black fireplace
a couch and a chair in a small living room
a living area with sofa, chair and a fireplace

Generated:
Baseline: a living room filled with furniture and a large window
Seq: a living room filled with furniture and a tv
Tok: a living room with a couch and a desk
Tok-Seq: a living room filled with furniture and a fireplace

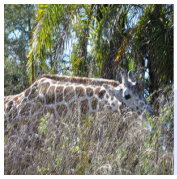
Ground truth:
a boy in a wet suit and a white and colored surfboar
a young child attempting to get onto a surfboard in the water
a young boy hanging on to a surfboard in the water
a young boy climbing into a surfboard in the water
a young man riding a surfboard in the ocean

Generated:
Baseline: a woman in the water with a surfboard
Seq: a woman is sitting on a surfboard in the water
Tok: a woman sitting on a surfboard in the water
Tok-Seq: a little boy that is standing in the water



Ground truth:
the airplane in the sky is doing tricks while spitting out smoke.
a plane flies through the air with fumes coming out the back
an airplane is letting off white smoke against a blue sky
a biplane leaves a smoke trail while doing a trick
a biplane flying upside down leaving a large vapor trail

Generated:
Baseline: a plane flying through a blue sky with clouds
Seq: an airplane flying in the sky with smoke coming from it
Tok: a small plane flying through a blue sky
Tok-Seq: a small plane flying through a blue sky



Ground truth:
a close up of a giraffe eating from the top of a tree
a giraffe bending over in tall grass by some trees
a tall giraffe eating leafy greens in a jungle
a standing giraffe in tall brush eating leaves
giraffe leaning very far over to sample leaves

Generated:
Baseline: a giraffe standing next to a tree in a field
Seq: a couple of giraffe standing next to each other
Tok: a giraffe standing in the middle of a forest
Tok-Seq: a giraffe eating leaves from a tree

Ground truth:
a purple flower hanging upside down from a green stem with green plantains attache
a plant with a purple flower and several unripe bananas
a flower hanging from a bunch or green bananas
a banana plant and a bunch of green bananas
a flower with some plants on its stem

Generated:
Baseline: a bunch of bananas hanging from a tree
Seq: a close up of a flower on a tree
Tok: a bunch of green bananas hanging from a tree
Tok-Seq: a green flower is hanging from a banana tree

Ground truth:
a man with a jacket posing to throw a frisbee outside.
a man is throwing a frisbee in a sandy area
a man about to throw a frisbee by the road
a man wearing a baseball hat tossing a frisbee
a man in a baseball cap throwing a frisbee.

Generated:
Baseline: a man is holding a frisbee in a field
Seq: a man in a hat is holding a tennis racket
Tok: a man in a hat is holding a frisbee
Tok-Seq: a man with a frisbee in his hand

Ground truth:
a woman looks apprehensive as she prepares to cut her own hair with scissors.
a woman that is holding her hair and a pair of scissors
a lady cutting her own hair with a pair of huge scissor
there is a woman holding scissors to her hair
a young woman is curling her long hair

Generated:
Baseline: a woman in a black shirt and a black tie
Seq: a woman in a black shirt and black hair
Tok: a woman in a black shirt and a black tie
Tok-Seq: a woman holding a pair of scissors in her hand

Ground truth:
black and white image of white umbrellas hanging below light
several umbrellas are hanging upside down from a lighting system
a bunch of ornate lanterns are hanging on the cielin
a warehouse filled with light fixtures in it
there are many lights hanging from this ceilin

Generated:
Baseline: a bunch of umbrellas hanging from a ceiling
Seq: a bunch of umbrellas that are in a room
Tok: a bunch of umbrellas that are in a building
Tok-Seq: a bunch of umbrellas hanging from a ceiling

Ground truth:
the man wearing a red tie is standing near an animal's plastic house
a young girl wearing a red tie is adjusting her glasses.
the man in the red tie is putting on his glasses
a person with glasses and a tie in a room
a woman wearing glasses a shirt and ti

Generated:
Baseline: a man in a tie is holding a cell phone
Seq: a woman standing in a room talking on a phone
Tok: a woman wearing a red shirt and a tie
Tok-Seq: a woman in a red shirt and red tie

Ground truth:
a bathroom with a red wall and poster in front of a toilet
there is an indoor toilet underneath a sign that says please flush.
this bathroom has red and white walls and a poster
a toilet with a poster above it in a bathroo
a sign is seen posted above a toilet

Generated:
Baseline: a toilet with a sign attached to it
Seq: a bathroom with a toilet and a sign
Tok: a toilet in a bathroom with a sign above it
Tok-Seq: a bathroom with a toilet and a sign

Ground truth:
a cat lies on the couch next to a computer while holding a red stuffed toy
a black and white cat laying on a couch holding a stuffed animal
a sleepy cat on its back playing with a stuffed anima
a black and white cat is cuddled with a red toy
a black and white cat with a red stuffed toy

Generated:
Baseline: a black and white cat laying on a blanket
Seq: a black and white cat laying on a bed
Tok: a black and white cat laying on a bed
Tok-Seq: a black and white cat laying on top of a bed

Ground truth:
a young man in a sweat shirt is standing on a wooden walkway
a person riding a skate board on a wooden sidewal
a man is riding a skateboard across a bridge
a boy riding his skateboard down the wooden deck.
a man skateboarding across a small wooden bridge.

Generated:
Baseline: a man riding a skateboard on top of a wooden rail
Seq: a man riding a skateboard down a metal rail
Tok: a man standing on top of a wooden bench
Tok-Seq: a man riding a skateboard on top of a wooden rail

Ground truth:
a couple is smiling while posing for a picture on a bed
a couple laying on a big bed in a bedroom
an image of a couple in bed on gold sheet
two people laying in a neatly made be
a couple is lying on the bed

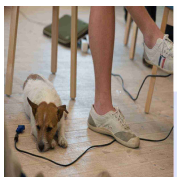
Generated:
Baseline: a woman sitting on a bed with a cat
Seq: a man sitting on a bed with a cat
Tok: a woman sitting on a bed in a bedroom
Tok-Seq: two people laying on a bed in a room

Ground truth:
two zebra's standing in a grassy field and one is eating grass
a zebra looking up as another grazes in a field
the zebras are grazing out in the field of grass.
a group of zebras stand together in a field
several zebras eating grass in a wildlife par

Generated:
Baseline: a couple of zebra standing on top of a grass covered field
Seq: a couple of zebra standing on top of a grass covered field
Tok: a couple of zebra standing next to each other on a field
Tok-Seq: a couple of zebras are standing in a field

Ground truth:
a man standing in front of a street vendor, a woman behind the counter
a man preparing food on a street cart next to a building
a man running a street stand in front of a yellow garag
a man and a woman stand at a portable vending cart
a woman eats a burger next to a street vendor

Generated:
Baseline: a man is standing next to a bicycle
Seq: a man on a bike with a man on the back
Tok: a couple of men standing next to each other
Tok-Seq: a man standing next to a woman in front of a store



Ground truth:
a wooden floor with a dog laying down next to a persons feet and next to
a grown and white dog on floor next to person's shoes
a dog laying on the floor next to a persons legs
a little dog laying under the table at someones fee
a small dog laying next to a person wearing sneaker

Generated:
Baseline: a dog sitting on the floor with a pair of shoes
Seq: a dog and a pair of shoes on a floor
Tok: a dog laying on the ground next to a person
Tok-Seq: a dog is laying on the floor next to a persons feet

Ground truth:
a large, twin engine airliner is slightly tilted to one side in the air
it's wondrous how that big airplane manages to stay up in the sky
an air canada plane making a left turn in the sky
a plane flying through the air during a sunny da
a plane on the air flying very hig

Generated:
Baseline: a large jetliner flying through a blue sky
Seq: a large jetliner flying through a blue sky
Tok: a large jetliner flying through a blue sky
Tok-Seq: an airplane flying in the air with a sky background

Ground truth:
a broken door frame showing a a bathroom with a cement floor and a broken back
an outside doorway to a restroom showing a destroyed wall and damaged floor
a bathroom with walls that are falling down and a toilet
a doorway leading into a delapitated wall in a room
a doorway looking into a demolished bathroo

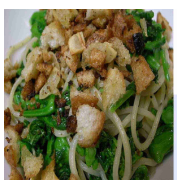
Generated:
Baseline: a dirty bathroom with a brick wall and a broken window
Seq: a door is open in a small room
Tok: a window that has a window in it
Tok-Seq: a dirty bathroom with a toilet and a window

Ground truth:
view from behind a tree of a lake and fall colored tress on the other side
a small boat near a wide river with a dense forest on the other sid
a snowy day in a colorful forest where leaves have already changed colors.
view from window across water with fall foliaged trees on other bank
a boat some water and some red and green tree

Generated:
Baseline: a view of a lake with a boat on it
Seq: a large body of water with a boat on it
Tok: a large body of water next to a forest
Tok-Seq: a view of a body of water with trees in the background

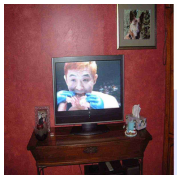
Ground truth:
a small glass of milk sitting next to a platter full of doughnuts
a cloth covered plate of donuts next to a glass of milk
the doughnuts are next to the glass of milk on the table
a couple of doughnuts sit next to a glass of milk
a savory snack of dounts with a glass of mil

Generated:
Baseline: a donut sitting on top of a wooden table
Seq: a close up of a doughnut on a plate
Tok: a donut and a drink on a table
Tok-Seq: a couple of donuts sitting on top of a table



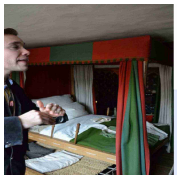
Ground truth:
bread crumbs sitting on top of a veggie plate with noodles.
a plate of food that includes broccoli, noodles and bread crumbs
a plate of pasta with greens and toasted bread pieces.
a pasta dish with bread crumbs and cooked green vegetabl
a white plate topped with salad and onions

Generated:
Baseline: a close up of a plate of food with broccoli
Seq: a plate of food with broccoli and meat
Tok: a plate of food with meat and vegetables
Tok-Seq: a close up of a plate of food



Ground truth:
a flat scree tv sitting on a wooden stand with an image of a ginger asian
a red headed man on a television in front of a red wall
a tv on a small table with pictures and figurines near b
an image of a living room setting with the tv o
a red painted wall is against a television

Generated:
Baseline: a television sitting on top of a wooden table
Seq: a flat screen tv mounted to a wall
Tok: a flat screen tv on a wooden table
Tok-Seq: a tv sitting on top of a wooden table

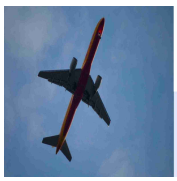


Ground truth:
a four poster bed covered with red and green curtains behind a ma
a man in his bedroom getting ready and a bed well mad
a man standing next to a bed while holding his hands together
a man standing near a bed with a red and green canopy
a man is standing near a canopy bed

Generated:
Baseline: a man sitting on a bed with a laptop
Seq: a man standing next to a bed covered in a blanket
Tok: a man sitting on a bed in a room
Tok-Seq: a man standing next to a bed with a canopy

Ground truth:
a man is on a horse and carriage by wells fargo
a chariot pulled by horses carrying people outside a building
a man and woman in a four horse drawn carriag
there are horses pulling a man and a cart
a horse carriage and horses moving along a street

Generated:
Baseline: a group of horses pulling a carriage down a street
Seq: a group of horses pulling a carriage down a street
Tok: a group of horses are pulling a carriage
Tok-Seq: a couple of horses pulling a carriage down a street



Ground truth:
a red and white plane flying under a blue sk
an orange, red and grey plane flying in the sky
a large jetliner flying through a cloudy blue sky
a airplane that is flying in the sky
the jet airplane flies across the blue sky.

Generated:
Baseline: an airplane flying in the sky with a sky background
Seq: a red and white plane flying in the sky
Tok: a plane flying in the air with a sky background
Tok-Seq: a red and white airplane flying in the sky

Ground truth:
a platter of raw vegetable crudites and dip sits on a table with other condiments.
a snack plate with dip carrots celery tomatoes pickles and cucumber on a tabl
a platter topped with sliced vegetables with ranch dip in the center
vegetable tray that contains carrots, cucumbers, peppers, tomatoes and celer
a vegetable platter with dip on a table

Generated:
Baseline: a plate of food including carrots and carrots
Seq: a table topped with a variety of vegetables
Tok: a plate of food with carrots and carrots
Tok-Seq: a table topped with lots of different vegetables

Ground truth:
a surfer standing on their board in relatively calm water
a woman riding a wave on top of a surfboard
a woman is on her surfboard in the water
she is well balanced on her new surfboard
a man that is surfing in some wate

Generated:
Baseline: a man on a surfboard in the water
Seq: a man on a surfboard in the water
Tok: a man riding a wave on top of a surfboard
Tok-Seq: a woman riding a wave on top of a surfboard

Ground truth:
a teenaged boy poses on the beach with his surfboard
a man stands on a beach with a surf board
a man poses with a surfboard on a beach
a man is standing next to a surfboard outsid
a guy on a beach holding a surf board

Generated:
Baseline: a man holding a surfboard on a beach
Seq: a man standing on a beach holding a surfboard
Tok: a man standing on a beach holding a surfboard
Tok-Seq: a man standing on a beach holding a surfboard

Ground truth:
a pizza that appears to be burned and has olives on it
a square shaped pizza with olives on it is over baked.
there is a pizza with cheese and olives on it
a rectangular pizza with cheese and olives sprinkled over it
a rectangular pizza with an egg on top

Generated:
Baseline: a close up of a pizza on a plate
Seq: a close up of a pizza on a plate
Tok: a pizza sitting on top of a white plate
Tok-Seq: a close up of a pizza on a table

Ground truth:
a couple of kids playing with a racquet in front of a the camera
a couple of girls sitting on a bench with tennis racquets
two kids holding tennis rackets while standing in a garage
two little girls and white uniforms holding up tennis rackets.
two young blonde girls sitting and holding tennis rackets

Generated:
Baseline: a woman and a little girl holding tennis rackets
Seq: a couple of kids standing next to each other
Tok: a woman and a child holding tennis rackets
Tok-Seq: two little girls sitting next to each other holding tennis rackets

Ground truth:
a man with safety equipment next to a fallen tree and red fire hydrant
the man is cutting down the trees around the red fire hydrant.
a worker standing next to a tree that's been chopped down.
a man standing by a tree and a fire hydrant
the man is cutting trees down outside.

Generated:
Baseline: a man standing next to a fire hydrant
Seq: a man sitting on a rock in the woods
Tok: a man sitting on a log in the woods
Tok-Seq: a man standing next to a red fire hydrant

C Neural machine translation

C.1 Experimental setups

WMT14 English-to-French We use the same experimental setting as Bahdanau et al. (2015): 12M paired sentences are used for training, 6,003 pairs for validation (news-test-2012 and news-test-2013) and 3,003 test pairs (news-test-2014). After tokenization, 30k most frequent tokens are selected for the model's vocabulary. We use an attentive encoder-decoder with a 2-layers bi-directionnal encoder of dimension $d = 2000$ and a single-layer decoder of dimension $d = 2000$ as well. We use batches of size 80 and train for 3 epochs with Adam (Kingma and Ba, 2015) starting with a learning rate of 2e-4. To generate translations we use beam search of size five.

IWSLT14 German-to-English We use the same settings as (Ranzato et al., 2016); the training set consists of 153k sentence pairs and 7k pairs are assigned to the validation and test sets. After tokenization and lower-casing, we remove sentences longer than 50 tokens. The English vocabulary has 22,822 words while the German has 32,009 words. We use an attentive encoder-decoder with a single bi-directionnal encoder and decoder of dimension $d = 128$. To generate translations we use beam search of size five. We use batches of size 32 and train for 40 epochs with Adam (Kingma and Ba, 2015) starting with a learning rate of 1e-3.

C.2 Examples

Source (en)	I think it's conceivable that these data are used for mutual benefit .
Target (fr)	J'estime qu'il est concevable que ces données soient utilisées dans leur intérêt mutuel .
MLE	Je pense qu'il est possible que ces données soient utilisées à des fins réciproques .
Tok-Seq	Je pense qu'il est possible que ces données soient utilisées pour le bénéfice mutuel .
Source (en)	However , given the ease with which their behaviour can be recorded , it will probably not be long before we understand why their tails sometimes go one way , sometimes the other .
Target (fr)	Toutefois , étant donné la facilité avec laquelle leurs comportements peuvent être enregistrés , il ne faudra sûrement pas longtemps avant que nous comprenions pourquoi leur queue bouge parfois d'un côté et parfois de l'autre .
MLE	Cependant , compte tenu de la facilité avec laquelle on peut enregistrer leur comportement , il ne sera probablement pas temps de comprendre pourquoi leurs contemporains vont parfois une façon , parfois l'autre .
Tok-Seq	Cependant , compte tenu de la facilité avec laquelle leur comportement peut être enregistré , il ne sera probablement pas long avant que nous ne comprenons la raison pour laquelle il arrive parfois que leurs agresseurs suivent un chemin , parfois l'autre .
Source (en)	The public will be able to enjoy the technical prowess of young skaters , some of whom , like Hyeres' young star , Lorenzo Palumbo , have already taken part in top-notch competitions .
Target (fr)	Le public pourra admirer les prouesses techniques de jeunes qui , pour certains , fréquentent déjà les compétitions au plus haut niveau , à l'instar du jeune prodige hyérois Lorenzo Palumbo .
MLE	Le public sera en mesure de profiter des connaissances techniques des jeunes garçons , dont certains , à l'instar de la jeune star américaine , Lorenzo , ont déjà participé à des compétitions de compétition .
Tok-Seq	Le public sera en mesure de profiter de la finesse technique des jeunes musiciens , dont certains , comme la jeune star de l'entreprise , Lorenzo , ont déjà pris part à des compétitions de gymnastique .

Table 7: WMT'14 English-to-French examples

Source (de)	sie repräsentieren teile der menschlichen vorstellungskraft , die in vergangene zeiten <UNK> . und für alle von uns , werden die träume dieser kinder , wie die träume unserer eigenen kinder teil der geographie der hoffnung .
Target (en)	they represent branches of the human imagination that go back to the dawn of time . and for all of us , the dreams of these children , like the dreams of our own children , become part of the naked geography of hope .
MLE	they represent parts of the human imagination that were blogging in the past time , and for all of us , the dreams of these children will be like the dreams of our own children part of hope .
Tok-Seq	and they represent parts of the human imagination that live in past times , and for all of us , the dreams of these children , like the dreams of our own children are part of hope .
Source (de)	und ja , vieles von dem , was heute gesagt wurde , berührt mich sehr , weil viele , viele schöne äußerungen dabei waren , die ich auch durchlebt habe .
Target (en)	and yes , a lot of what is said today really moves me , because many , many nice statements were made , which i also was part of .
MLE	and yes , a lot of what 's been said to me today , i got very , very , very , very beautiful expressions that i used to live through .
Tok-Seq	and yes , a lot of what 's been told today is very , very much , because many , lots of beautiful statements that i 've been through .
Source (de)	noch besser , er wurde in <UNK> nach den angeblich höchsten standards der nachhaltigkeit gezüchtet .
Target (en)	even better , it was <UNK> to the supposed highest standards of sustainability .
MLE	even better , he has been bred in love with the highest highest standards of sustainability .
Tok-Seq	even better , he was raised in terms of dignity , the highest standards of sustainability .

Table 8: IWSLT'14 German-to-English examples