IDENTIFYING AND CONTROLLING IMPORTANT NEURONS IN NEURAL MACHINE TRANSLATION

Paper Reading

Longyue Wang

Introduction

Motivation

- not clear fully distributed or individual neurons
- non-trivial linguistic information
- limitations in previous work: 1) whole vector representation (computer vision); 2) external supervision and costly annotation

Hyperthesis

Different models learn similar properties, and do not require any costly external supervision

Introduction

Questions:

- individual neurons to MT?
- individual neurons to linguistic interpretation?
- how to control neurons for improvement?

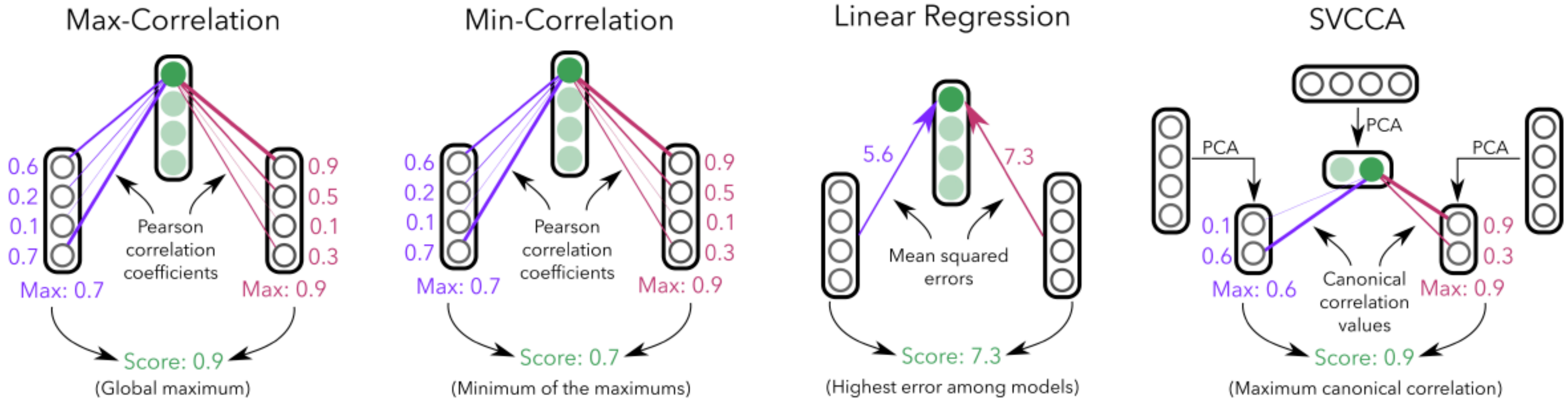
Inspiration (computer vision)

- Li et al. (2016): models v.s. properties
- Bau et al. (2017) individual neurons

Method

Unsupervised Correlation:

- Maximum correlation $\texttt{MaxCorr}(\mathrm{x}_i^m) = \max_{j,m' \neq m} |\rho(\mathrm{x}_i^m, \mathrm{x}_j^{m'})|$
- Minimum correlation $\texttt{MinCorr}(\mathrm{x}_i^m) = \min_{m' \neq m} \max_{j} |\rho(\mathrm{x}_i^m, \mathrm{x}_j^{m'})|$
- regression ranking: linear regression mean squared error
- SVCCA: PCA



Method

Verifying Detected Neurons:

- **Erasing**: masking ranked neurons according to correlation
- **Supervision**: expected conditional variance of neuron activations conditioned on some properties
- **Visualisation**: activations of neuron

Experiment

Data:

- UN corpus: multiple parallel
- 5 languages: AR, ZH, RU, SP, EN*
- 3*500K: 18 models

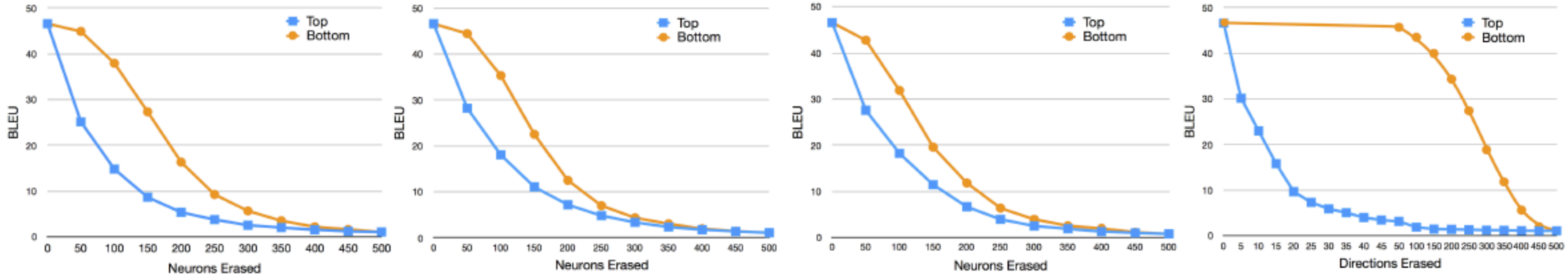
Models

- 2-layer LSTM encoder-decoder
- no BPE but char-CNN for morphology
- Transformer for future work

Results

Erasing:

- Neurons ranked higher by our methods have a larger impact on translation quality
- Top SVCCA directions capture very important information in the model



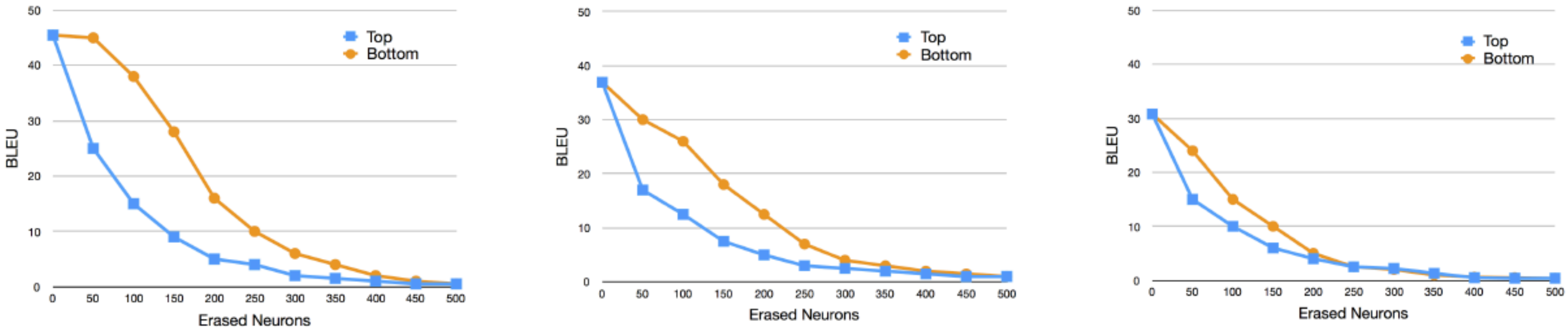
(a) MaxCorr (b) MinCorr (c) LinReg (d) SVCCA

Figure 2: Erasing neurons (or SVCCA directions) from the top and bottom of the list of most important neurons (directions) ranked by different unsupervised methods, in an English-Spanish model.

Results

Erasing:

- In all cases, erasing from the top hurts performance more than erasing from the bottom



(a) English-Spanish

(b) English-French

(c) English-Chinese

Figure 3: Erasing neurons from the top or bottom of the `MaxCorr` ranking in three language pairs.

Results

Evaluation Top Neurons:

- What kind of information is captured by the neurons ranked highly by each of our ranking methods?
- The percent of variance in neuron activation that is eliminated by conditioning on position in the sentence

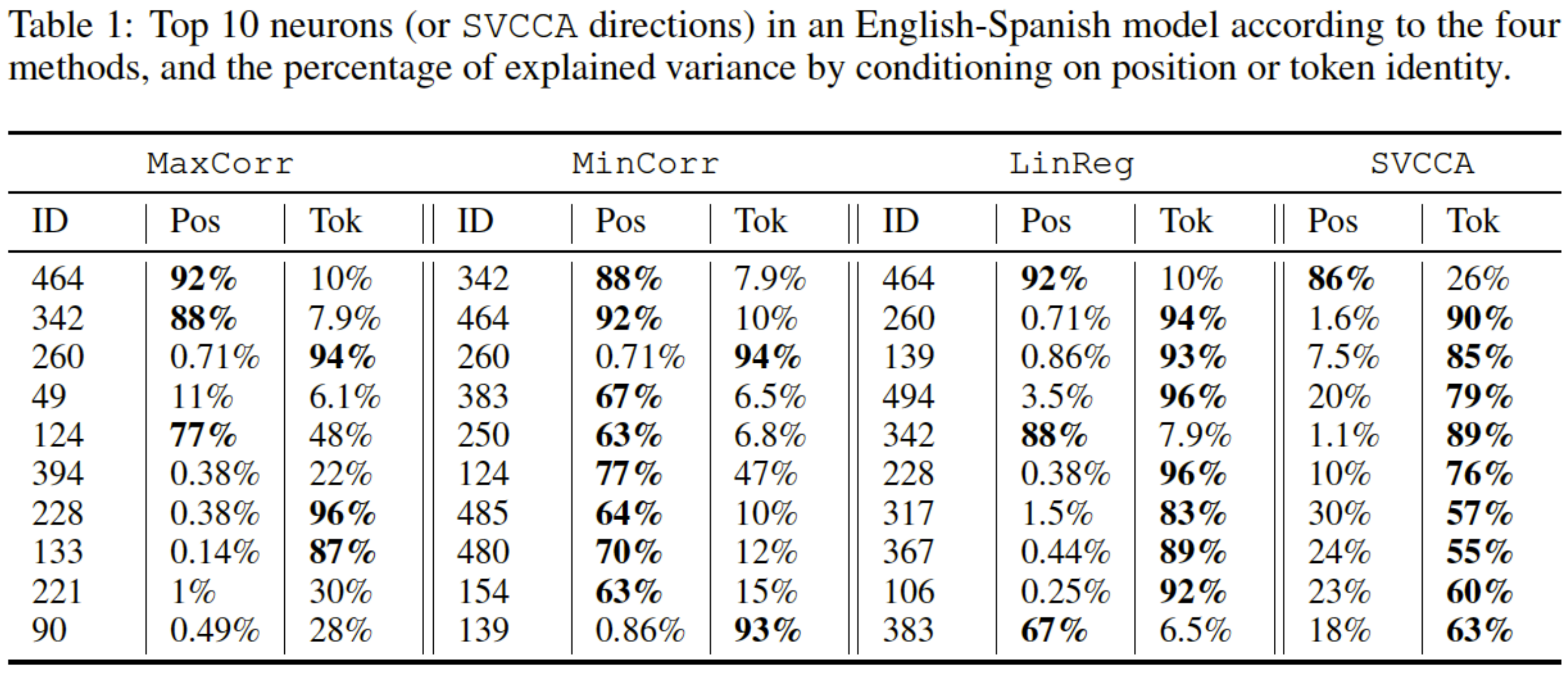


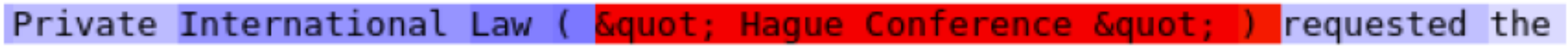
Table 1: Top 10 neurons (or SVCCA directions) in an English-Spanish model according to the four methods, and the percentage of explained variance by conditioning on position or token identity.

MaxCorr			MinCorr			LinReg			SVCCA	
ID	Pos	Tok	ID	Pos	Tok	ID	Pos	Tok	Pos	Tok
464	**92%**	10%	342	**88%**	7.9%	464	**92%**	10%	**86%**	26%
342	**88%**	7.9%	464	**92%**	10%	260	0.71%	**94%**	1.6%	**90%**
260	0.71%	**94%**	260	0.71%	**94%**	139	0.86%	**93%**	7.5%	**85%**
49	11%	6.1%	383	**67%**	6.5%	494	3.5%	**96%**	20%	**79%**
124	**77%**	48%	250	**63%**	6.8%	342	**88%**	7.9%	1.1%	**89%**
394	0.38%	22%	124	**77%**	47%	228	0.38%	**96%**	10%	**76%**
228	0.38%	**96%**	485	**64%**	10%	317	1.5%	**83%**	30%	**57%**
133	0.14%	**87%**	480	**70%**	12%	367	0.44%	**89%**	24%	**55%**
221	1%	30%	154	**63%**	15%	106	0.25%	**92%**	23%	**60%**
90	0.49%	28%	139	0.86%	**93%**	383	**67%**	6.5%	18%	**63%**

Results

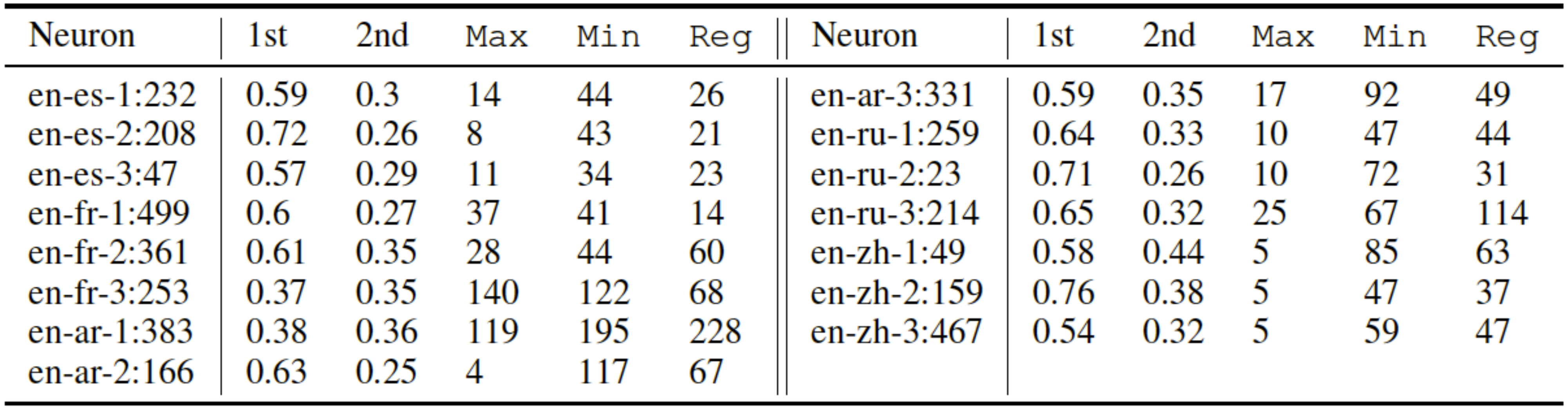
Linguistically Interpretable Neurons:

- Parentheses: neurons that detect parentheses were ranked highly in most models by the MaxCorr method



```
Private International Law ( &quot; Hague Conference &quot; ) requested the
```

Table 2: F_1 scores of the top two neurons from each network for detecting tokens inside parentheses, and the ranks of the top neuron according to our intrinsic unsupervised methods.

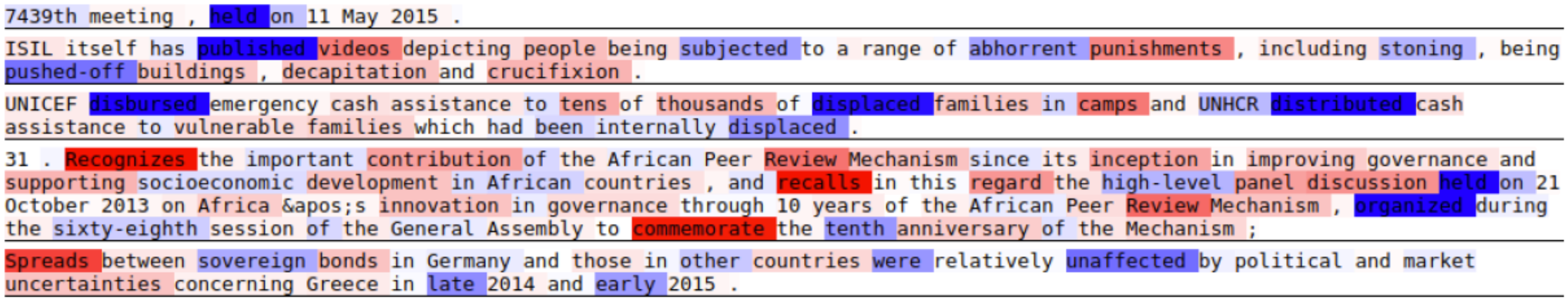


Neuron	1st	2nd	Max	Min	Reg	Neuron	1st	2nd	Max	Min	Reg
en-es-1:232	0.59	0.3	14	44	26	en-ar-3:331	0.59	0.35	17	92	49
en-es-2:208	0.72	0.26	8	43	21	en-ru-1:259	0.64	0.33	10	47	44
en-es-3:47	0.57	0.29	11	34	23	en-ru-2:23	0.71	0.26	10	72	31
en-fr-1:499	0.6	0.27	37	41	14	en-ru-3:214	0.65	0.32	25	67	114
en-fr-2:361	0.61	0.35	28	44	60	en-zh-1:49	0.58	0.44	5	85	63
en-fr-3:253	0.37	0.35	140	122	68	en-zh-2:159	0.76	0.38	5	47	37
en-ar-1:383	0.38	0.36	119	195	228	en-zh-3:467	0.54	0.32	5	59	47
en-ar-2:166	0.63	0.25	4	117	67						

Results

Linguistically Interpretable Neurons:

- Tense: we annotated the test data for verb tense (with Spacy) and trained a GMM model to predict tense from neuron activations
- This suggests that tense emerges in a “real” NMT model, but not in an auto-encoder that only learns to copy.



7439th meeting , held on 11 May 2015 .

ISIL itself has published videos depicting people being subjected to a range of abhorrent punishments , including stoning , being pushed-off buildings , decapitation and crucifixion .

UNICEF disbursed emergency cash assistance to tens of thousands of displaced families in camps and UNHCR distributed cash assistance to vulnerable families which had been internally displaced .

31 . Recognizes the important contribution of the African Peer Review Mechanism since its inception in improving governance and supporting socioeconomic development in African countries , and recalls in this regard the high-level panel discussion held on 21 October 2013 on Africa &apos;s innovation in governance through 10 years of the African Peer Review Mechanism , organized during the sixty-eighth session of the General Assembly to commemorate the tenth anniversary of the Mechanism ;

Spreads between sovereign bonds in Germany and those in other countries were relatively unaffected by political and market uncertainties concerning Greece in late 2014 and early 2015 .

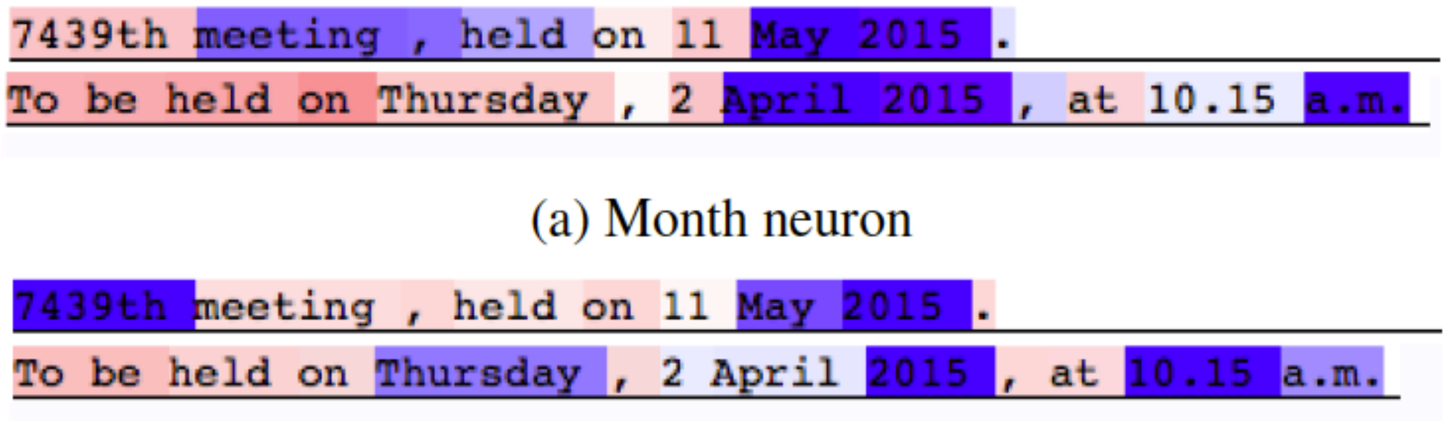
Table 3: Strongest correlations in all models relative to a tense neuron in an English-Arabic model.

Arabic	0.66, 0.57	French	-0.69, -0.58, -0.48	Chinese	-0.51, -0.30, -0.18
Spanish	0.56, 0.36, 0.22	Russian	-0.50, -0.39, -0.29	English	-0.33, -0.19, -0.03

Results

Linguistically Interpretable Neurons:

- Others: we found neurons that activate on numbers, dates, adjectives, plural nouns, auxiliary verbs, and more.


7439th meeting , held on 11 May 2015 .
To be held on Thursday , 2 April 2015 , at 10.15 a.m.
7439th meeting , held on 11 May 2015 .
To be held on Thursday , 2 April 2015 , at 10.15 a.m.

(a) Month neuron

(b) Approximate "year" neuron

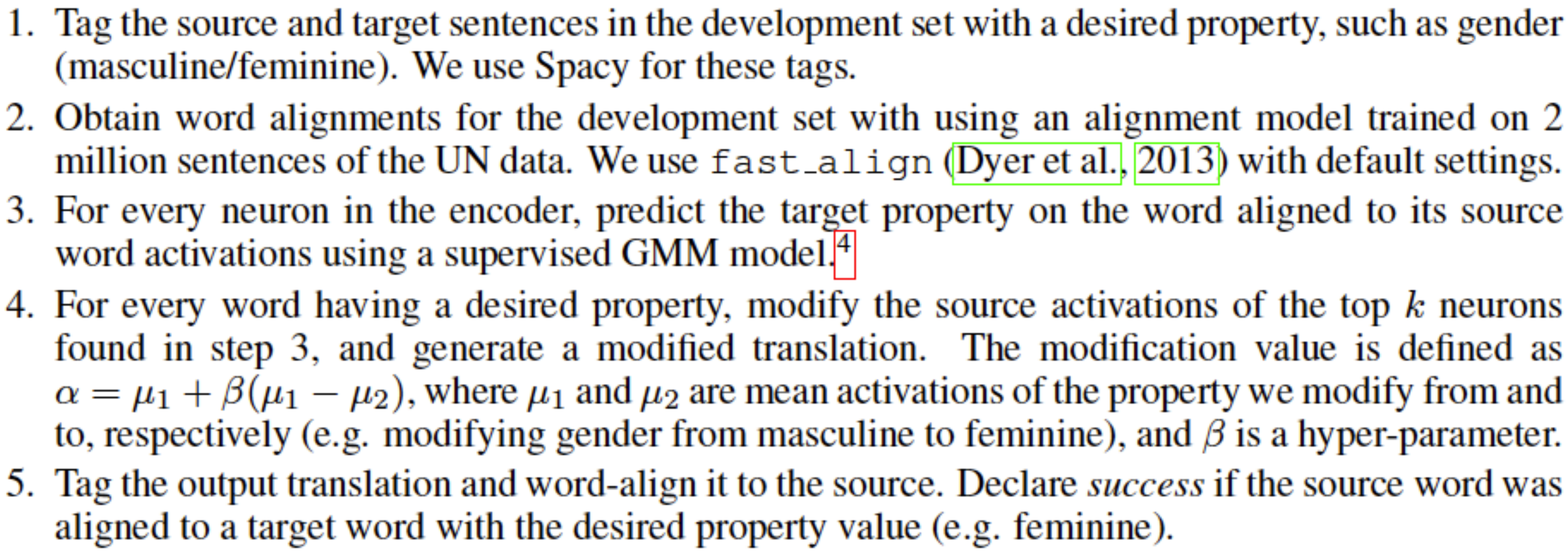
Figure 6: Neurons capturing dates and numbers.

Results

Controlling Translation:

(1) a. o bir doctor
b. he is a doctor

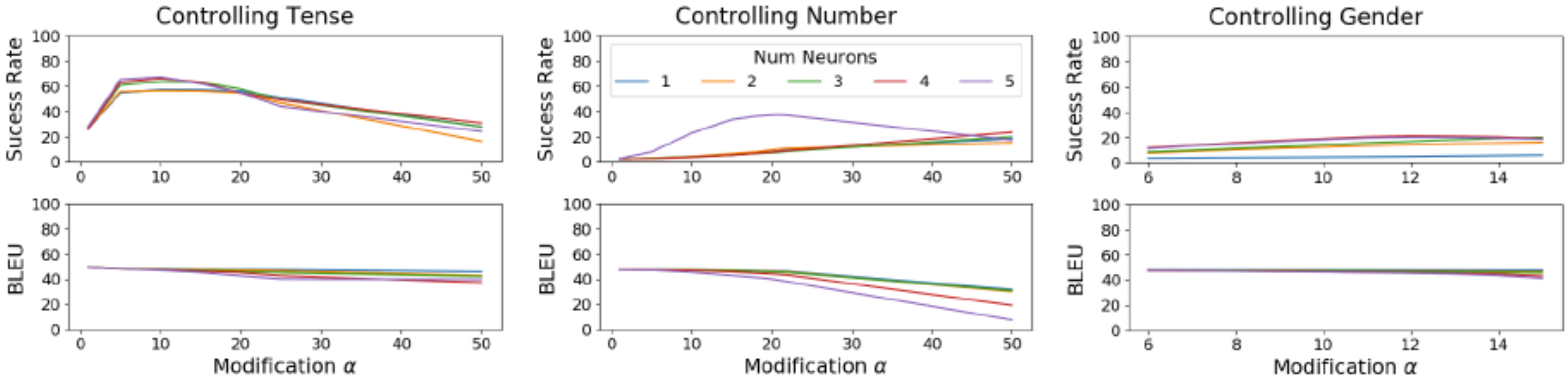
(2) a. o bir hemşire
b. she is a nurse



1. Tag the source and target sentences in the development set with a desired property, such as gender (masculine/feminine). We use Spacy for these tags.
2. Obtain word alignments for the development set with using an alignment model trained on 2 million sentences of the UN data. We use `fast_align` (Dyer et al., 2013) with default settings.
3. For every neuron in the encoder, predict the target property on the word aligned to its source word activations using a supervised GMM model.[4]
4. For every word having a desired property, modify the source activations of the top k neurons found in step 3, and generate a modified translation. The modification value is defined as $\alpha = \mu_1 + \beta(\mu_1 - \mu_2)$, where μ_1 and μ_2 are mean activations of the property we modify from and to, respectively (e.g. modifying gender from masculine to feminine), and β is a hyper-parameter.
5. Tag the output translation and word-align it to the source. Declare *success* if the source word was aligned to a target word with the desired property value (e.g. feminine).

Results

Controlling Translation:



(a) Tense (b) Number (c) Gender

Figure 4: Success rates and BLEU scores for controlling NMT by modifying neuron activations.

Table 4: Examples for controlling translation by modifying activations of different neurons on the *italicized* source words. α = modification value (–, no modification).

(a) Controlling number when translating "The interested *parties*" to Spanish.

α	Translation	Num	α	Translation	Num
-1	abiertas particulares	pl.	0.125	La parte interesada	sing.
-0.5	Observaciones interesadas	pl.	0.25	Cuestion interesada	sing.
-0.25, -0.125, 0	Las partes interesadas	pl.	0.5, 1	Gran útil	sing.

(b) Controlling gender when translating "The interested *parties*" (left) and "*Questions* relating to information" (right) to Spanish.

α	Translation	Gen	α	Translation	Gen
-0.5, -0.25	Los partidos interados	ms.	-1	Temas relativos a la información	ms.
0, 0.25	Las partes interesadas	fm.	-0.5, 0, 0.5	Cuestiones relativas a la información	fm.

(c) Controlling tense when translating "The committee *supported* the efforts of the authorities".

	α	Translation	Tense
Arabic	–/+10	وأيدت\وتؤيد اللجنة {جهود\الجهود التي تبذلها} السلطات	past/present
French	–/-20	Le Comité a appuyé/appuie les efforts des autorités	past/present
Spanish	–/-3/0	El Comité apoyó/apoyaba/apoya los esfuerzos de las autoridades	past/impf./present
Russian	–/-1	Комитет поддержал/поддерживает усилия властей	past/present
Chinese	–/-50	委员会 支持 当局 的 努力 / 委员会 正在 支持 当局 的 努力	untensed/present