Universiteit Leiden

Netherlands Organisation for Scientific Research

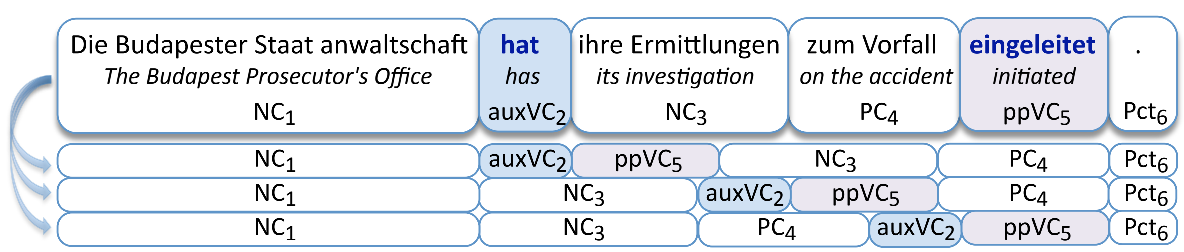
Unveiling the Linguistic Weaknesses of Neural Machine Translation

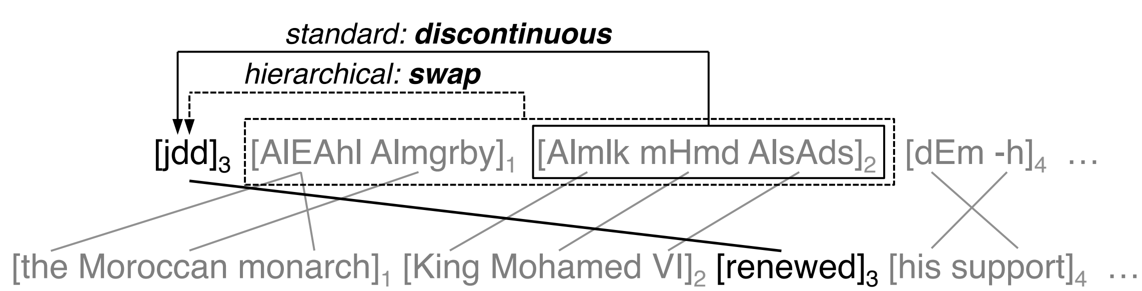
Arianna Bisazza

my first encounter with neural machine translation

Back in 2014…

I had been working on word reordering models for five years



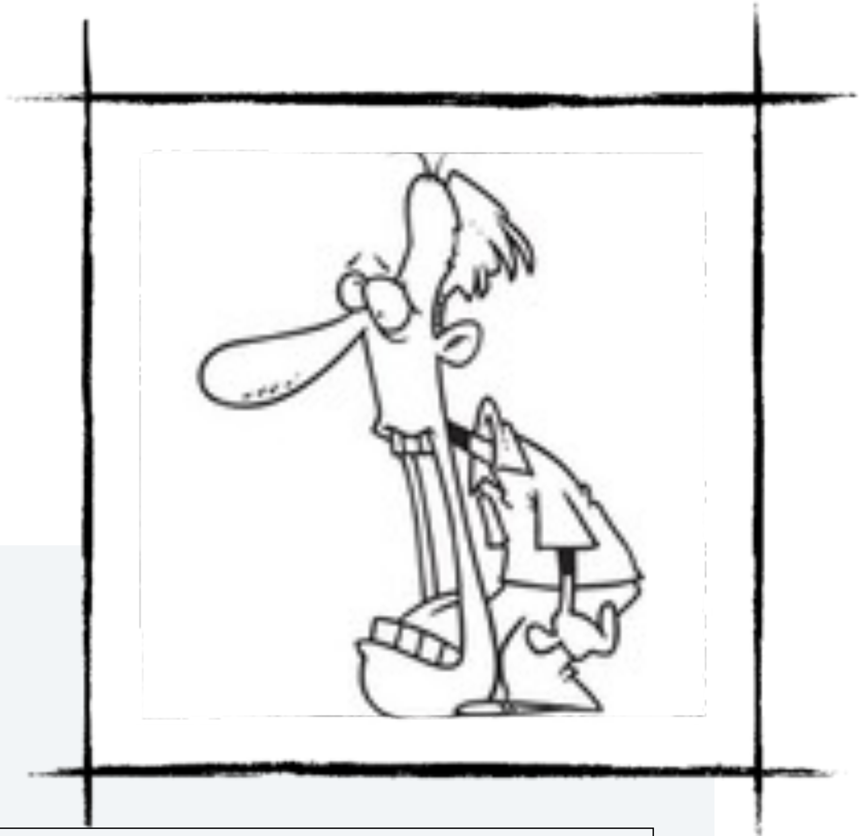


Reordering models	**References**	**Model type**	**Reordering step classification**	**Features**
Phrase orientation models (POM): Example: P(*orient=discontinuous-left* | *next-phrase-pair=[jdd]-[renewed]*)				
lexicalized (hierarchical) phrase orientation model	Tillmann 2004; Koehn & al. 2005; Nagata & al. 2006; Galley & Manning 2008	gener.	monotonic, swap, discontinuous (left or right)	source/target phrases
phrase orientation maxent classifier	Zens & Ney 2006	discr.		source/target words or word clusters
sparse phrase orientation features	Cherry 2013	discr.		
Jump models (JM): Example: P(*jump=−5* | *from=AlsAds, to=jdd*)				
inbound/outbound/pairwise lexicalized distortion	Al-Onaizan & Papineni 2006	gener.	jump length	source words
inbound/outbound length-bin classifier	Green & al. 2010	discr.	jump length based (9 length bins)	source words, POS, position; sent. length
Source decoding sequence models (SDSM): Example: P(*next-word=jdd* | *prev-translated-words=AlEahil Almlk mHmd AlsAds*)				
reordered source n-gram	Feng & al. 2010a	gener.	—	source words (9-gram context)
source word-after-word	Bisazza & Federico 2013; Goto & al. 2013	discr.	—	source words, POS; source context's words and POS
Operation sequence models (OSM): Example: P(*next-operation=generate[jdd,renewed]* | *prev-operations=generate[AlsAds,VI] jumpBack[1]*)				
translation/reordering operation n-gram	Durrani & al. 2011; Durrani & al. 2013; Durrani & al. 2014	gener.	insertGap, jumpBack, jumpForward	source/target words, POS or word clusters; prev. $n-1$ operations

I was integrating a neural component for word translation prediction into SMT

Back in 2014…

Montreal's first NMT online demo:



Type text here:

The Budapest Prosecutor's Office has initiated an investigation on the accident.

Translation:

Die Budapester Staatsanwaltschaft hat ihre Ermittlungen zum Vorfall eingeleitet.

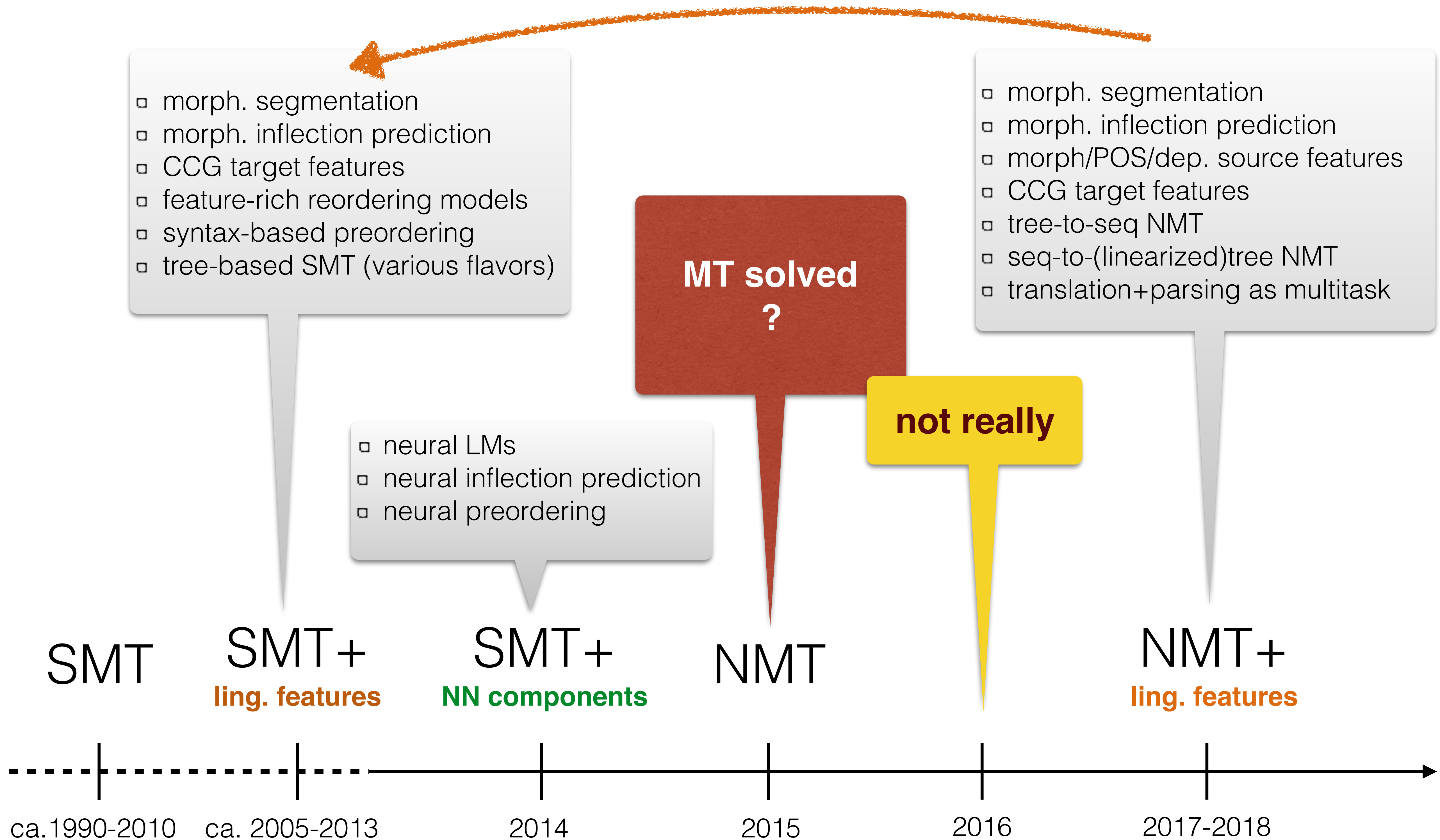
New research direction

- My interests suddenly switched to **discovering the strengths and weaknesses** of neural seq(-to-seq) models
- In 2016 published **first error analysis** of NMT vs SMT output post-editing

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	

[Bentivogli,Bisazza,Cettolo,Federico. EMNLP'16]

History repeats itself



Let’s take a step back

Do we know where we are going?

This time we’re dealing with a really black box

- In pre-neural SMT we knew what could *not* work by model limitations (e.g. clearly flawed independence assumptions)
- Neural models have the potential to learn *anything*, but *do* they in practice?

Research should aim at:

- understanding the role played by linguistic structure in seq(-to-seq) models
- more systematic ways to know which linguistic phenomena are(n’t) captured [→ model interpretability]

Today's talk

(1) What makes recurrent NNs work so well for language modeling?

(2) How important is recurrency for capturing hierarchical structure?

(3) Do NMT models learn to extract linguistic features from raw data and exploit them in any explicable way?

Part 1:

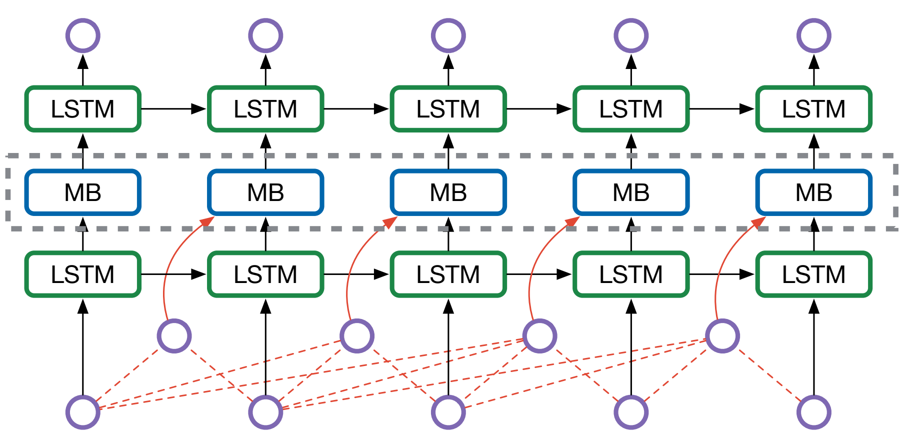
What makes recurrent NNs work so well for language modeling?

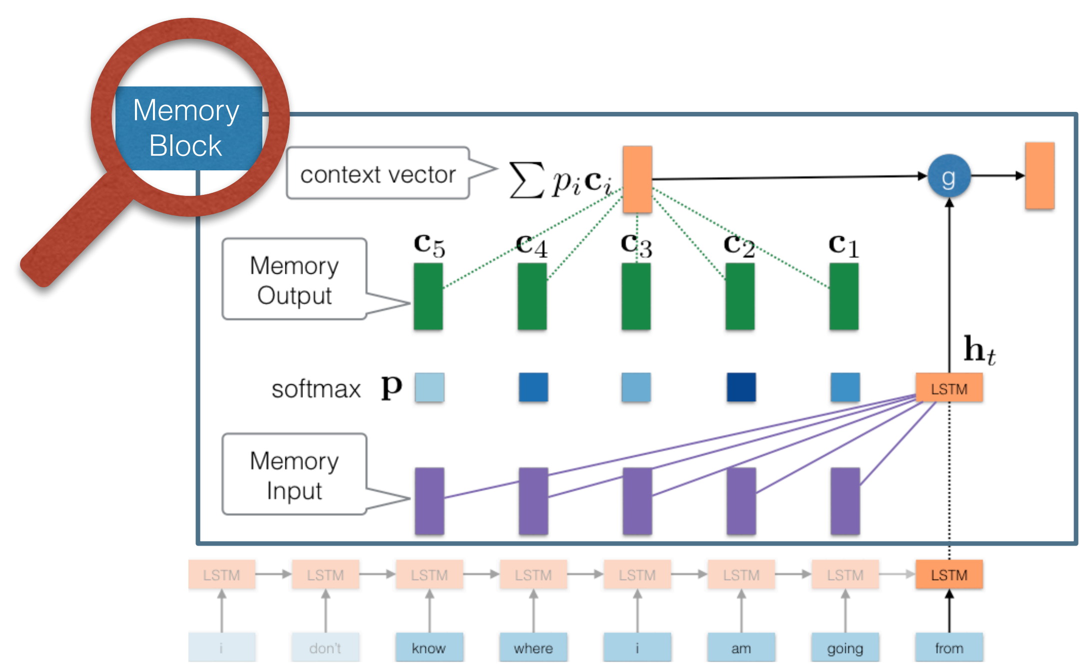
First Insights into the Workings of RNNs

Our first hypothesis: a great command of language structure (grammar)

How to find that out?

- Augment an LSTM language model with a memory block (precursor to self-attention)
- Read out the weights of attention over the last *n* words
- Test on language modeling: essential subtask of machine translation and other seq-to-seq tasks



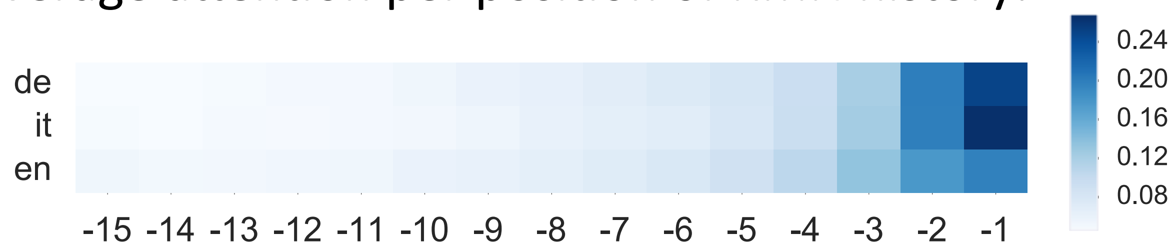


[Tran,Bisazza,Monz. NAACL'16]

First Insights into the Workings of RNNs

Attention visualization on 100 word samples (DE):

Average attention per position of RMN history:



Long-dependency examples:

wie wirksam die daraus resultierende strategie sein wird , hängt daher von der genauigkeit dieser annahmen

Gloss: *how effective the from-that resulting strategy be will, depends therefore on the accuracy of-these measures*

Translation: *how effective the resulting strategy will be, therefore, depends on the accuracy of these measures*

ab (-1.8)
und (-2.1)
, (-2.5)
. (-2.7)
von (-2.8)

… die lage versetzen werden , eine schlüsselrolle bei der eindämmung der regionalen ambitionen chinas zu

Gloss: … *the position place will, a key-role in the curbing of-the regional ambitions China's to*

Translation: *…which will put him in a position to play a key role in curbing the regional ambitions of China*

spielen (-1.9)
gewinnen (-3.0)
finden (-3.4)
haben (-3.4)
schaffen (-3.4)

[Tran,Bisazza,Monz. NAACL'16]

First Insights into the Workings of RNNs

✔ Lexical co-occurrences

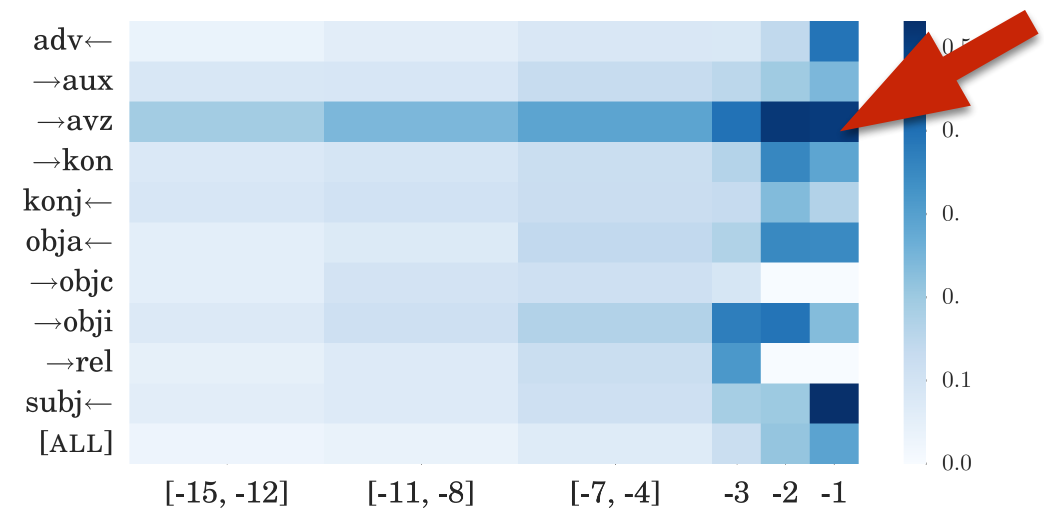
Frequent pairs of *mostAttendedWord-predictedWord* with distance >6 words:

German	English Trans
findet *statt*	takes *place*
kehrte *zuruck*	came *back*
fragen *antworten*	questions *answers*
kämpfen *gegen*	fight *against*
bleibt *erhalten*	remains *intact*
verantwortung *übernimmt*	*takes* responsibility

Italian	English Trans
sinistra *destra*	left *right*
latitudine *longitudine*	latitude *longitude*
collegata *tramite*	connected *through*
sposò *figli*	got-married *children*
insignito *titolo*	awarded *title*

? Syntactic dependencies

- only to a limited extent
- mostly separable verbs (in German)



Later work [Linzen & al. 2016] confirmed and explained our findings:
LSTM captures long syntactic dependencies *iff* explicit supervision is used

[Tran,Bisazza,Monz. NAACL'16]

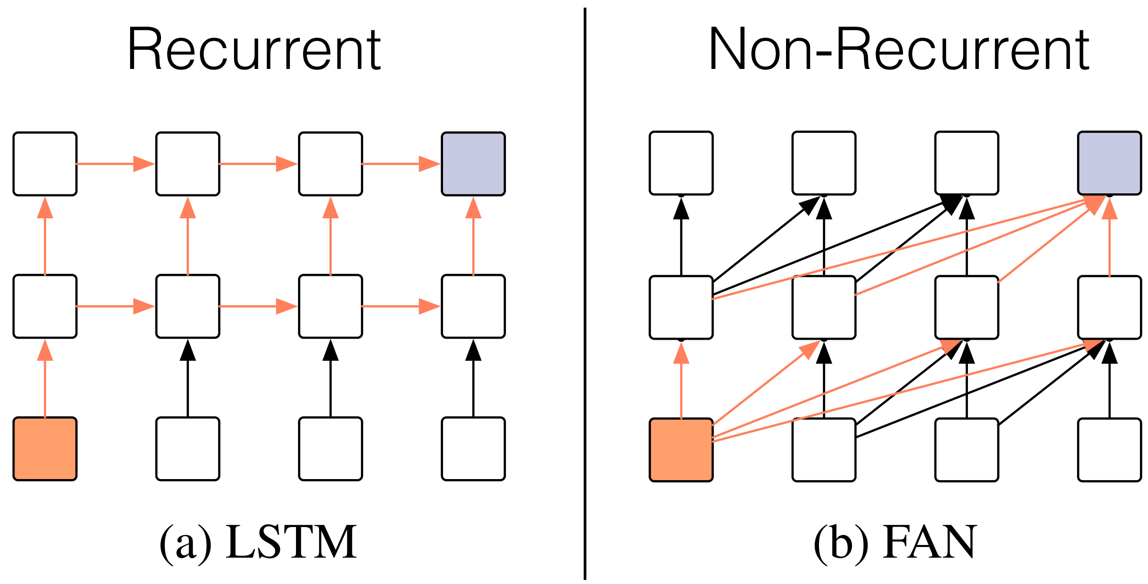
Part 2:

How important is recurrency for capturing hierarchical structure?

The Importance of Being Recurrent

Recently a family of non-recurrent models show competitive performance on seq-to-seq modeling, esp. machine translation:

- **CNNs** (Convolutional Neural Networks) [Gehring & al. 2017]
- **FANs** (Fully Attentional Networks) [Vaswani & al. 2017]



- ✓ Lower complexity
- ✓ Much more parallelizable
- ✓ Shorter path among input positions

But does this kind of models capture hierarchical structure?

Capturing hierarchical structure is necessary to truly understand, process and translate language

[Tran,Bisazza,Monz. arXiv 2018]

The Importance of Being Recurrent

We choose two tasks where capturing hierarchical structure is strictly required:

- **subject-verb agreement** [Linzen & al. 2016]:

The **keys** to the <u>cabinet</u> **are** on the table.

- Predict verb number: **are/is** ?

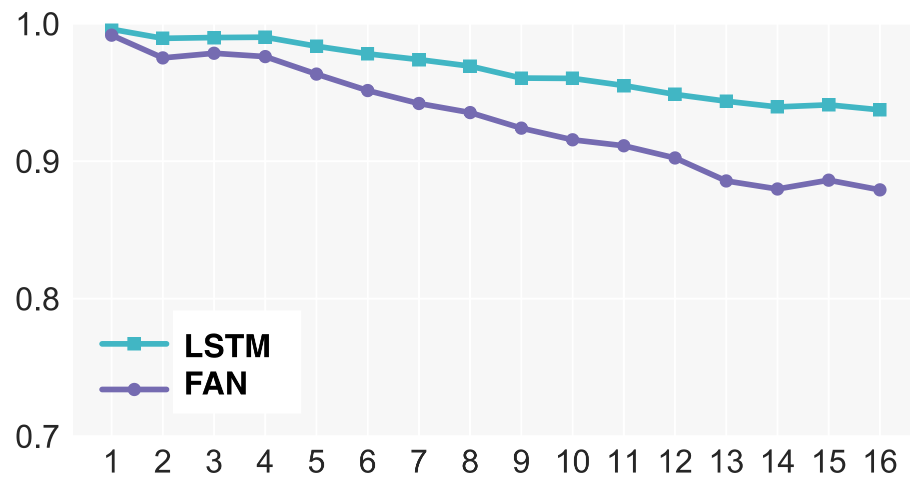
- **logical inference** [Bowman & al. 2015]:

$$(d (\text{or } f)) \sqsupset (f (\text{and } a))$$
$$(d (\text{and } (c (\text{or } d)))) \# (\text{not } f)$$
$$(\text{not } (d (\text{or } (f (\text{or } c))))) \sqsubset (\text{not } (c (\text{and } (\text{not } d))))$$

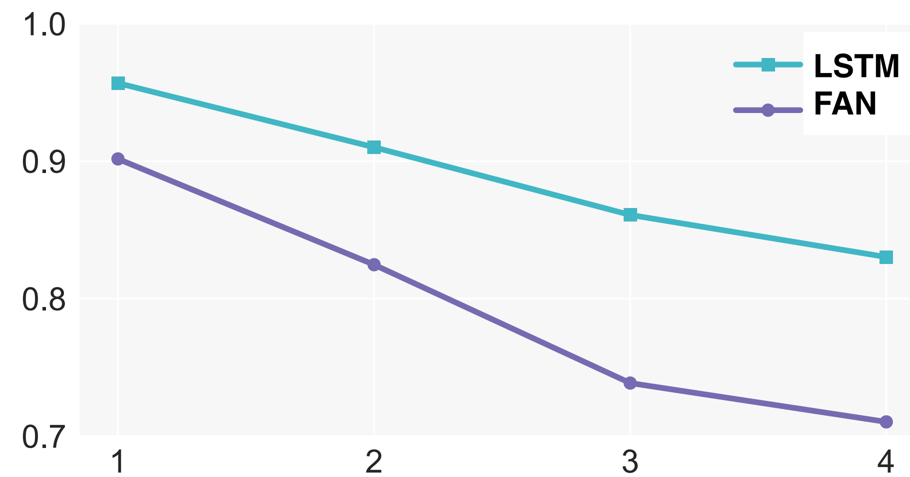
- Predict 1 of 7 logical relations
- Artificial data

[Tran,Bisazza,Monz. arXiv 2018]

Results(1) Subject-Verb Agreement



(c) Number prediction, breakdown by distance

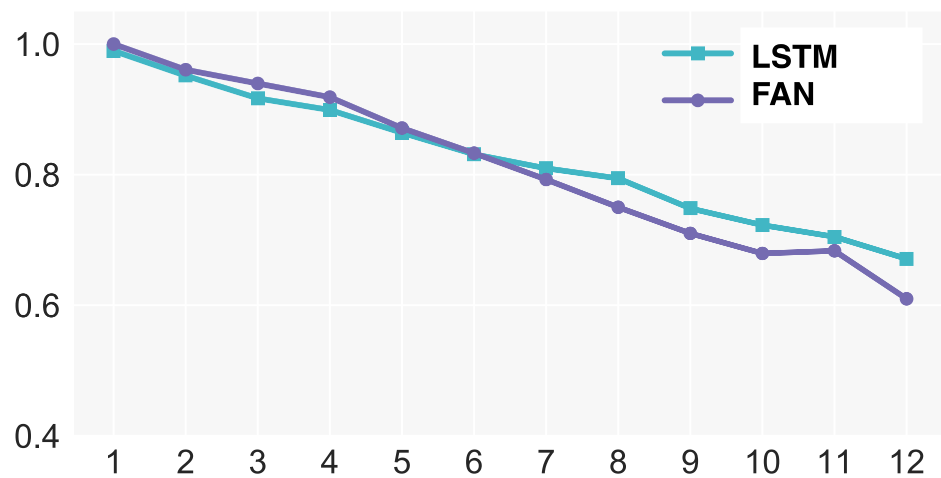


(d) Number prediction, breakdown by # attractors

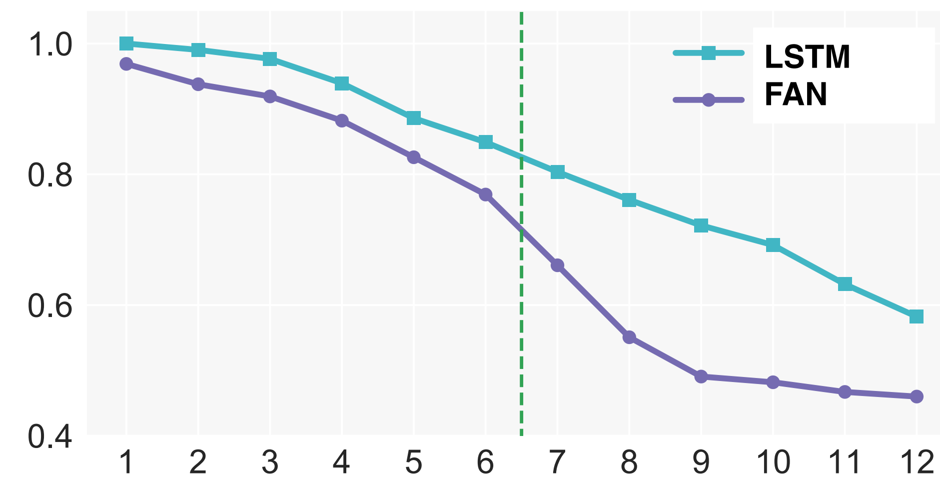
- Both models achieve high performance
- LSTM slightly but consistently better and more robust to task difficulty
- (FAN has lower perplexity though)

[Tran,Bisazza,Monz. arXiv 2018]

Results(2) Logical Inference



(b) $n \leq 12$



(a) $n \leq 6$

- Similar performance when trained on whole data
- LSTM much better than FAN when only trained on short sequences (generalization power)

[Tran,Bisazza,Monz. arXiv 2018]

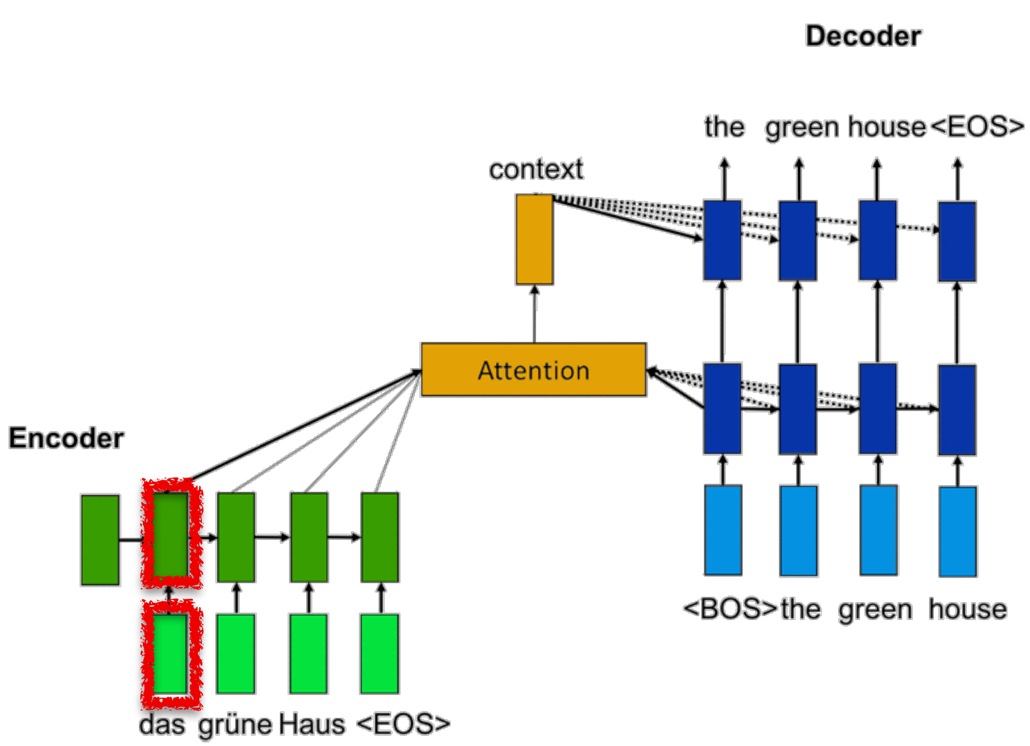
Part 3:

Do NMT models extract linguistic features from raw data and exploit them in explicable ways?

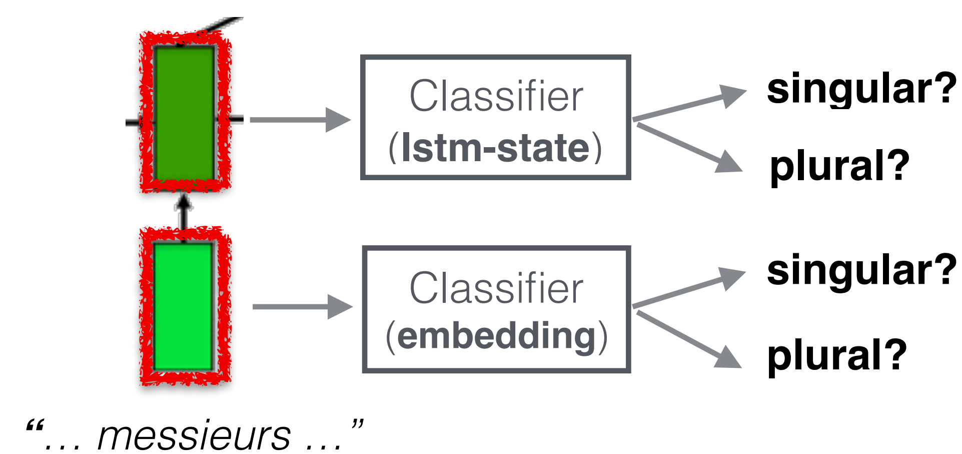
Morphological features in NMT embeddings

Potential: understand if injecting linguistic knowledge into machine translation (e.g. via supervised annotation) is a promising direction

- Specifically, we look at morphology on the source side
- Build on and extend first analysis by [Belinkov & al. 2017]
- Method: Train linguistic classifiers on word representations produced by NMT encoders



https://aws.amazon.com/blogs/machine-learning



[Bisazza, Tump. *In Preparation*]

Experimental Setup

NMT:

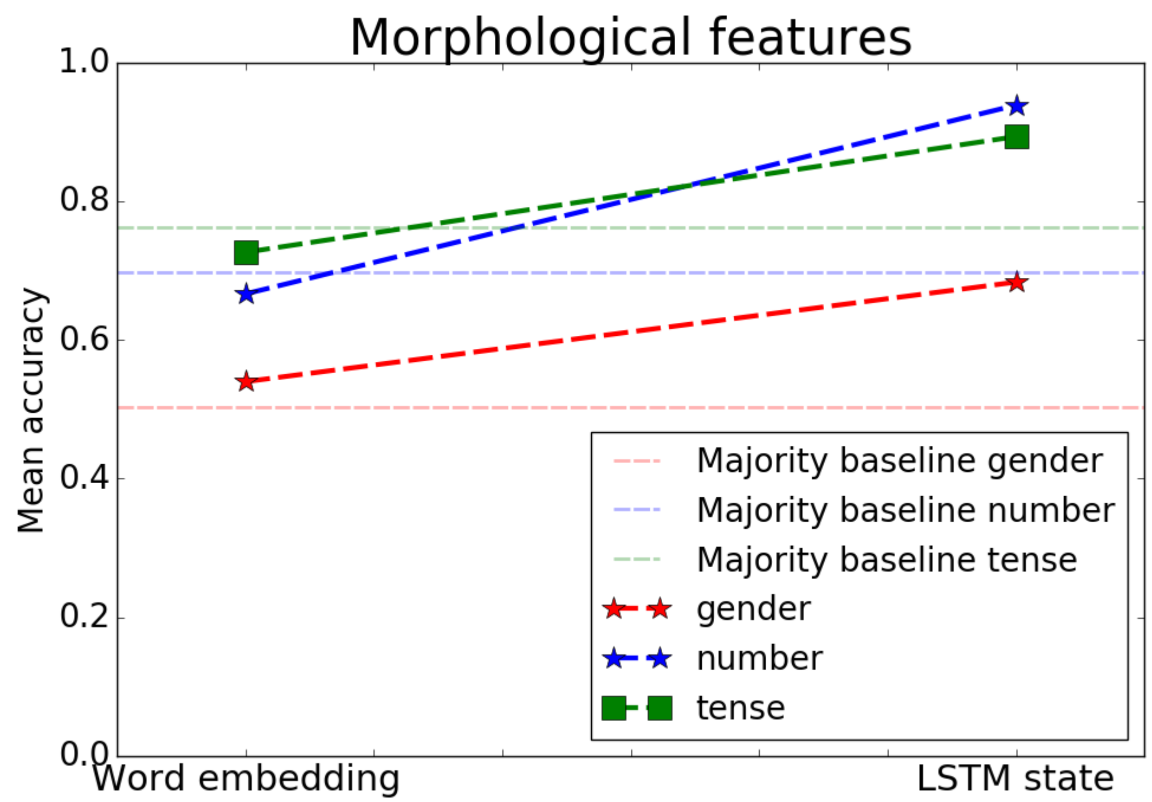
- Language pairs: French→Italian/German/English
- Always analyze source-side (French) vectors
- NMT model: word-level, 3-layer LSTM, |h|=1000, |dict| = 30K
- BLEU: 32.6 (FR-IT), 25.4 (FR-DE), 39.4 (FR-EN)

Classifiers:

- Linear classifiers
- Labels from morphological lexicon
- No vocabulary overlap between training and test (essential to avoid overfitting)

[Bisazza,Tump. *In Preparation*]

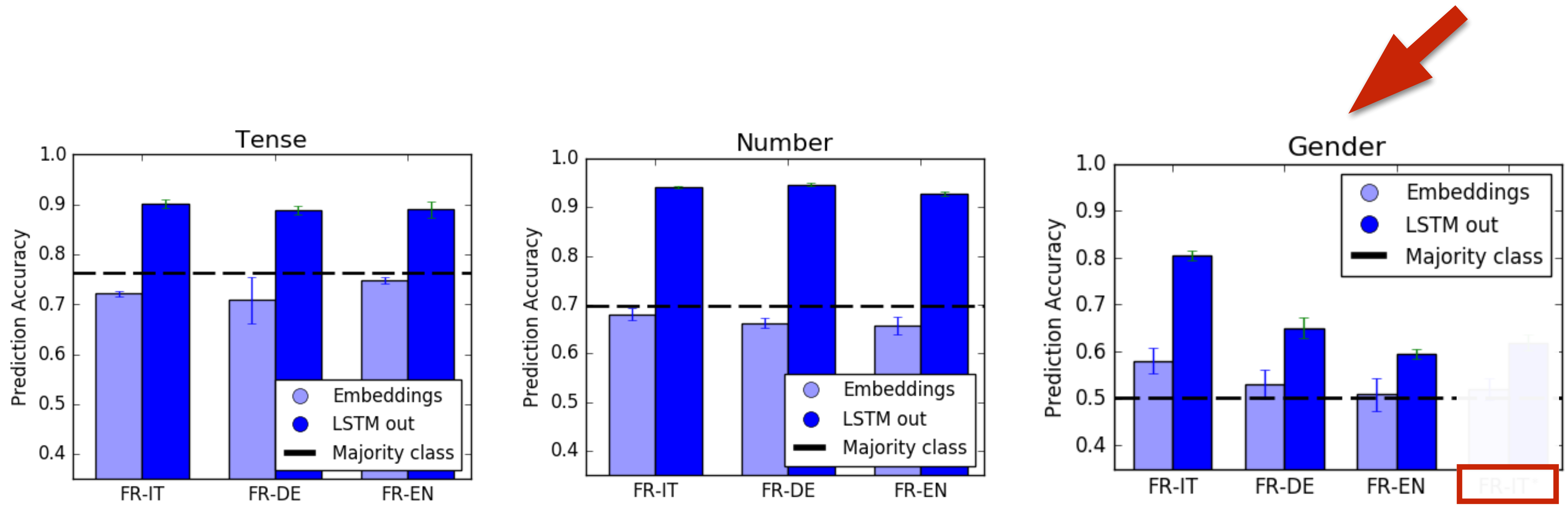
Results for All Target Languages



- Source morphological features only encoded *in-context*, not as word type properties (→ morph. information not stored in the lexicon!)
- Semantic features (*number*, *tense*) encoded much better than purely grammatical features (*gender*)

[Bisazza,Tump. *In Preparation*]

Impact of Target Language



- Morphology is *not* learned better when translating into morphologically poorer English (diff. from previous findings)
- Impact of target language only visible on *gender*
- FR-IT*: much lower gender accuracy when removing target gender marking

All suggest that morphological features are **only learned** when **directly transferrable** to target

[Bisazza,Tump. *In Preparation*]

to conclude

Summary

- RNNs clearly capture lexical co-occurrences, but syntax only to a limited extent (unless provided with explicit supervision)
- Recurrency is important to properly capture hierarchical structure
- NMT models learn and exploit linguistic features only when directly transferable to target language

RNNs are powerful models of language and have no rivals when it comes to capturing implicit structure.
Still, their command of syntax remains imperfect and poorly interpretable.

What’s next

We need more interpretable models:

- to deliver reliable technology
- to detect limitations and address them

Mainly a responsibility of the Machine Learning community …?

… NLP’ers also need to ask the right questions:

- what makes a model *interpretable* in the language domain?
- less quantitative, more qualitative evaluation: an age shift

  → design challenge sets requiring specific language competence to be solved [Linzen & al. ’16][Sennrich’17][Burlot & Yvon ’17]
- many more phenomena and languages remain to be covered

  → (semi-)automate challenge set creation, e.g. using existing parsers

  → explore general benefits of combining specific supervision objectives

Thanks for your attention

Arianna Bisazza @ Leiden University

References

- Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In Proceedings of EMNLP 2016, Austin, USA, 2016
- Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo and Marcello Federico. Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. In Computer Speech & Language, 49:52-70, 2018.
- Ke Tran, Arianna Bisazza and Christof Monz. Recurrent Memory Networks for Language Modeling. In Proceedings of NAACL 2016, San Diego, USA, 2016.
- Ke Tran, Arianna Bisazza and Christof Monz. *The Importance of Being Recurrent for Modeling Hierarchical Structure. arXiv:1803.03585,* 2018*.*
- Arianna Bisazza and Clara Tump. *On the Role of Morphology in Neural Machine Translation. [In preparation]*