Universiteit Leiden

Netherlands Organisation for Scientific Research

Is Morphology Captured by Neural Machine Translation?

Arianna Bisazza

Motivation

Myth or fact?

- Neural language/translation models don't need feature engineering
- Continuous word representations capture fine-grained lexical properties
- Recurrent NNs capture long-range dependencies

History repeats itself

- morph. segmentation
- morph. inflection prediction
- CCG target features
- feature-rich reordering models
- syntax-based preordering
- tree-based SMT (various flavors)

MT s ved
?

- morph. segmentation
- morph. inflection prediction
- morph/POS/dep. source features
- CCG target features
- tree-to-seq NMT
- seq-to-(linearized)tree NMT
- translation+parsing as multitask

- neural LMs
- neural inflection prediction
- neural preordering

unsupervised solutions to the rare word problem

SMT
ca.1990-2010

SMT+
ling. features
ca. 2005-2013

SMT+
NN components
2014

NMT
2015

subword/char-level
NMT
2016

NMT+
ling. features
2017

What's new

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?
- Harder to make explaining hypotheses, harder to test them

BLEU-like evaluation metrics becoming obsolete:

- More correct translations not matching the reference

Today's talk

I - Overview of recent work investigating what linguistic phenomena are (not) captured by NMT models

II - On-going experiments on morphology features contained in NMT's internal representations

Part I

Many approaches

- (Semi-)manual **error analysis**
  - [Bentivogli & al. ’16] detected reordering as a major strength of NMT vs PBMT

Analyzing human post-edited data

[Bentivogli,Bisazza,Cettolo,Federico. EMNLP 2016]

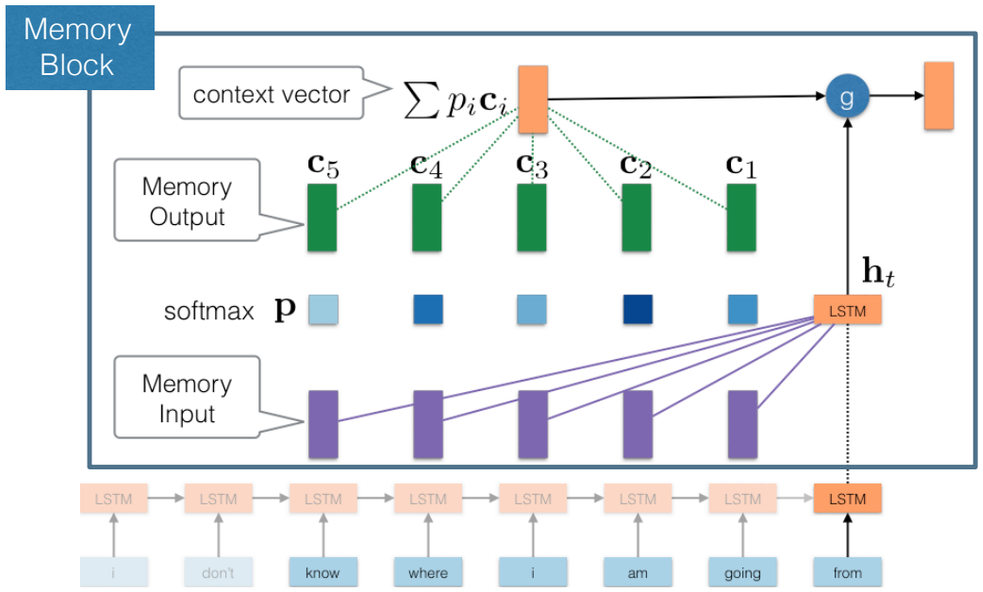
*Auxiliary-main verb construction [aux:*V*]:*			
	SRC	in this experiment , individuals **were shown** hundreds of hours of YouTube videos	
(a)	HPB	in diesem Experiment , Individuen **gezeigt wurden** Hunderte von Stunden YouTube-Videos	✗
	PE	in diesem Experiment **wurden** Individuen Hunderte von Stunden 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 ...	
(b)	PBSY	... wenn Trainer und Manager und Eigentümer **betrachten** diese Information Streaming ...	✗
	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**	
(c)	SPB	so wie viele von uns , ich habe in ein paar Schränke **in meinem Leben** gelebt	✗
	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	
(d)	HPB	aber ich kam schlielich zu dem Schluss , dass nur aus systematischen Gründen **nicht** funktionieren	✓
	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	

Many approaches

- (Semi-)manual **error analysis**
  - [Bentivogli & al. ’16] detected reordering as a major strength of NMT vs PBMT
- Provide **linguistic annotation** to the model, see if quality improves
  - mixed results; typically best on small data
- **Test suites**: design tasks needing linguistic competence to be solved
  - contrastive sentence pairs [Linzen & al. ’16][Sennrich’17][Burlot & Yvon ’17]
- Examine NMT’s representations by **transfer learning**
  - parser/morph.classifier trained on NMT encoded vectors [Shi & al.’16][Belinkov & al.’17][*this talk*]
- Modify the **model** to be more **interpretable**
  - (self-)attention, memory networks, representation erasure

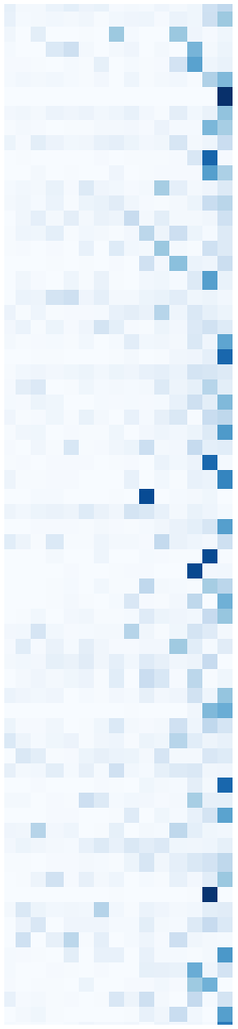
Recurrent Memory Network

[Tran,Bisazza,Monz. NAACL 2016]

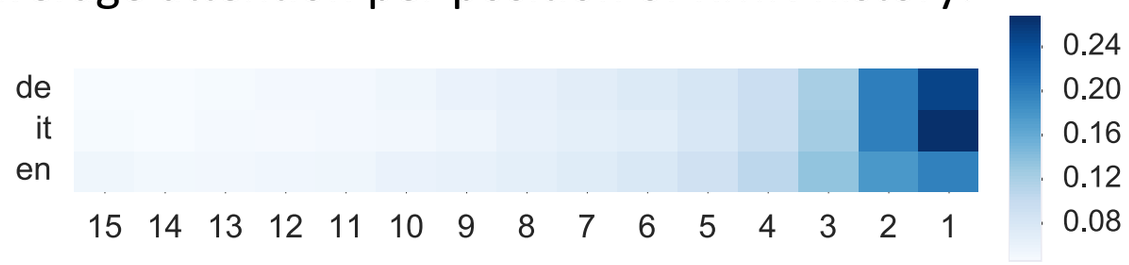


wie wirksam die daraus resultierende strategie sein wird , hängt daher von der genauigkeit dieser annahmen ab

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*

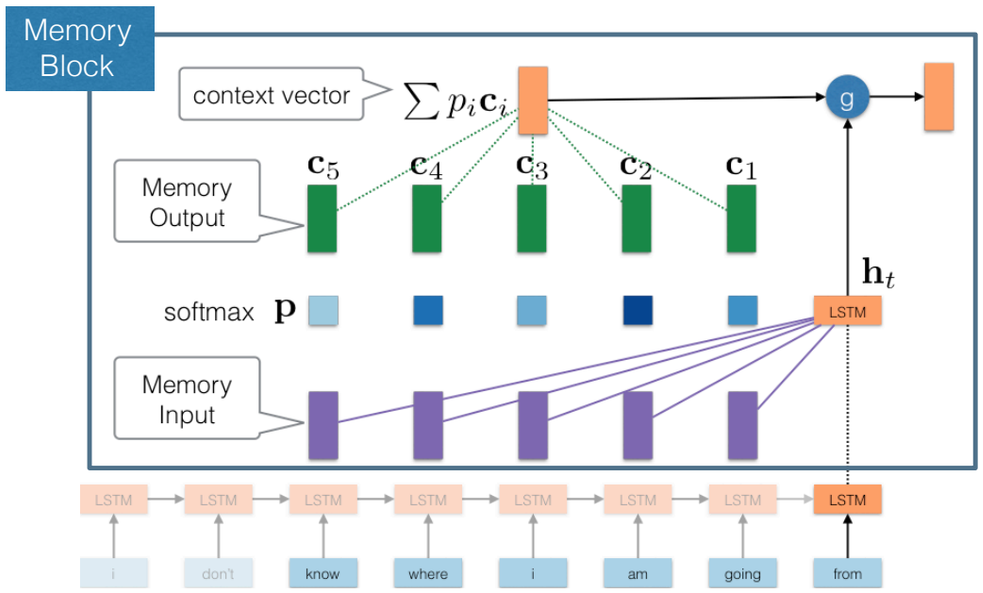


Average attention per position of RMN history:



Attention visualization on 100 word samples (de)

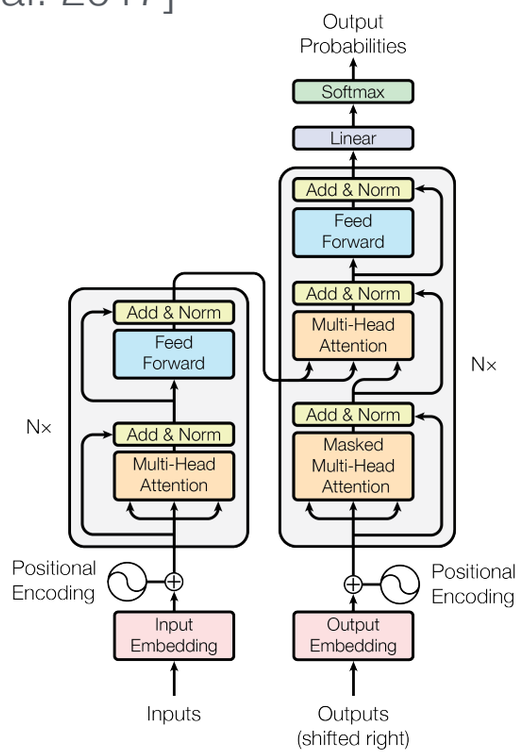
Recurrency *vs.* Attention

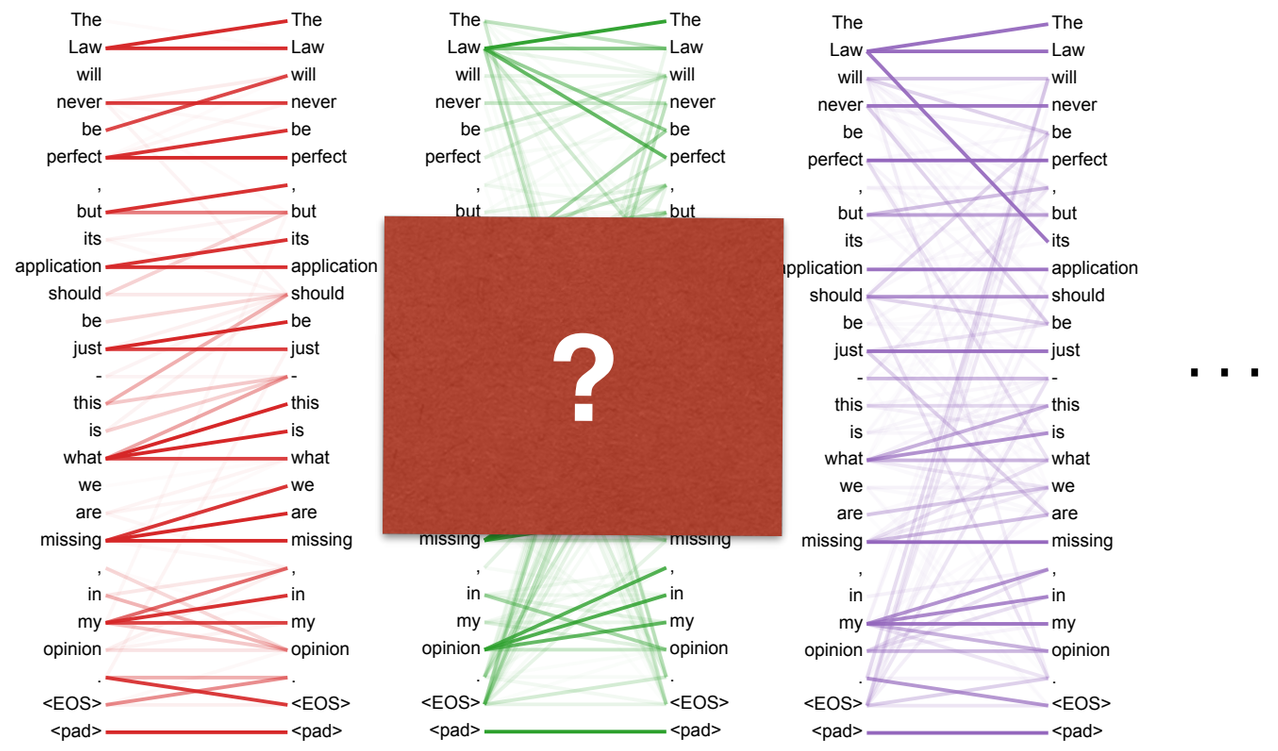


wie wirksam die daraus resultierende strategie sein wird , hängt daher von der genauigkeit dieser annahmen ab

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*

[Vaswani & al. 2017]





Mixed findings

Positive evidence:

- NMT spots subj-verb & det-noun agreement errors with near-human accuracy [Sennrich'17]
- Parse tree extracted from NMT sentence vector with high accuracy [Shi&al.'16]

Negative/conflicting evidence:

- Memory-augmented LM doesn't give particular attention to agreement triggers [Tran & al.'16]
- Supervised morph. segmentation always better than unsupervised for LM perplexity [Vania & Lopez '17]
- BPE-based NMT *vs.* morphology-aware NMT: no clear winner [Burlot&Yvon'17]
- Source morphology better captured when target language is 'easier' even if morphologically poor [Belinkov & al.'16]
- Character-level NMT captures morphological features better than word-level NMT [Belinkov & al.'16] but is worse at agreement [Sennrich'17]

Part II

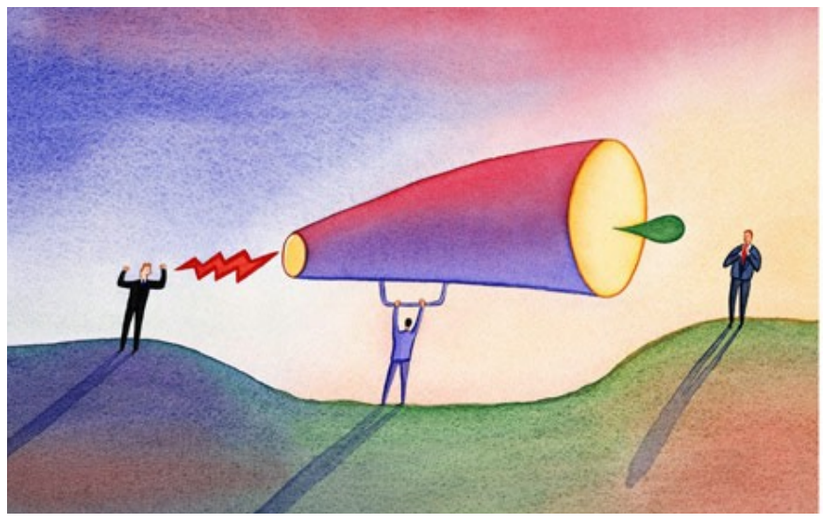
Morphological features in NMT embeddings

[Bisazza&Tump. *On-going*]

- Research Question: Does the model learn linguistic features to understand the source text and render it in the target language as we expect a human would?

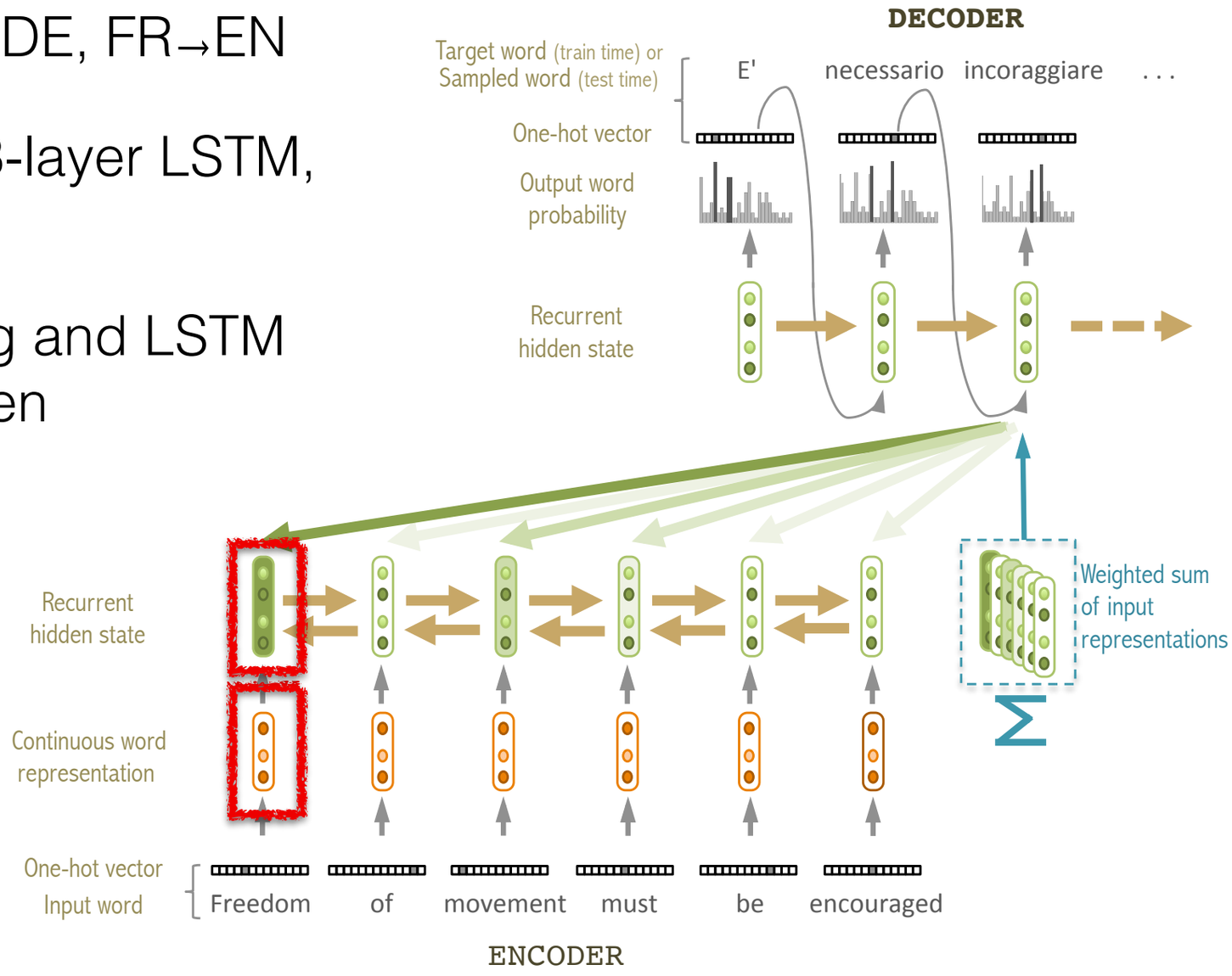
- Approach: Transfer learning

- Data: French to Italian/German/English



Method (1)

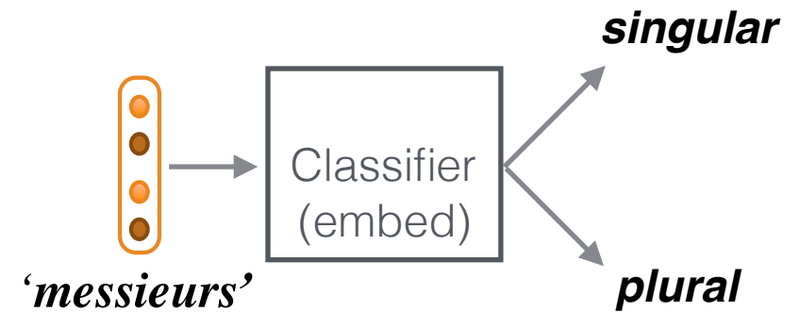
- Train NMT on FR→IT, FR→DE, FR→EN
- NMT model: word-level, 3-layer LSTM, |h|=1000, |dict| = 30K
- Take out word embedding and LSTM state for each French token

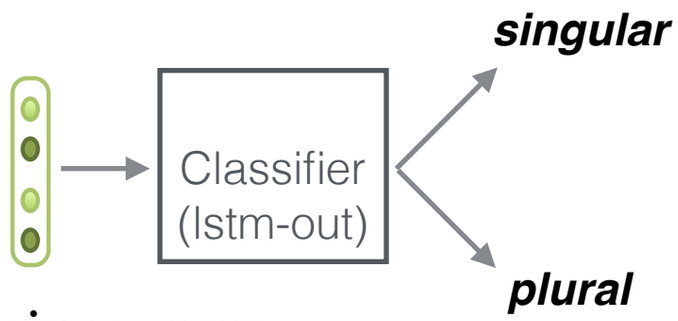


*BLEU scores: 32.6 (FR-IT), 25.4 (FR-DE), 39.4 (FR-EN)

Method (2)

- Build a linear classifier
- Labels from morphological lexicon
- Randomly split vocabulary into 2 non-overlapping parts (train/test)*
- Repeat 5 times and average

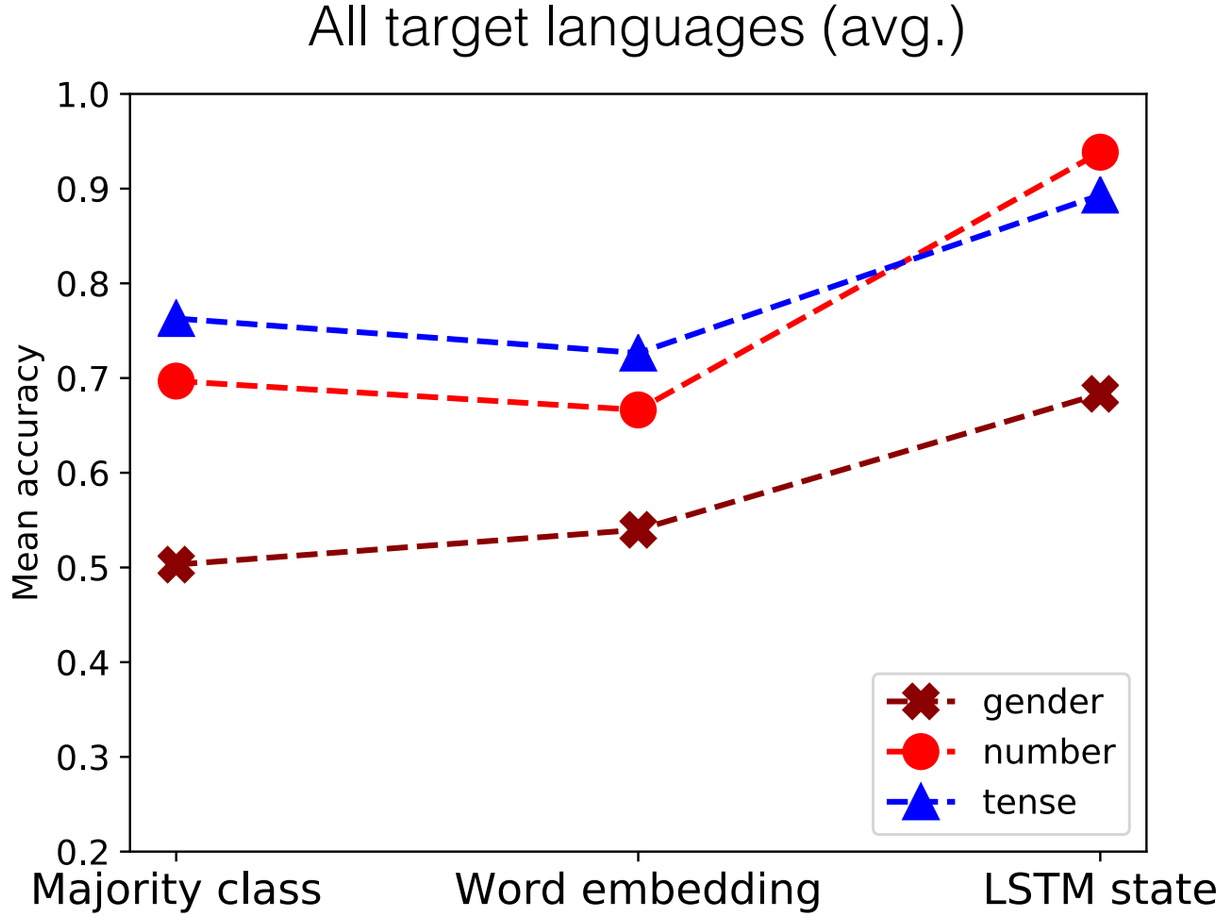




*Essential step to avoid major overfitting!

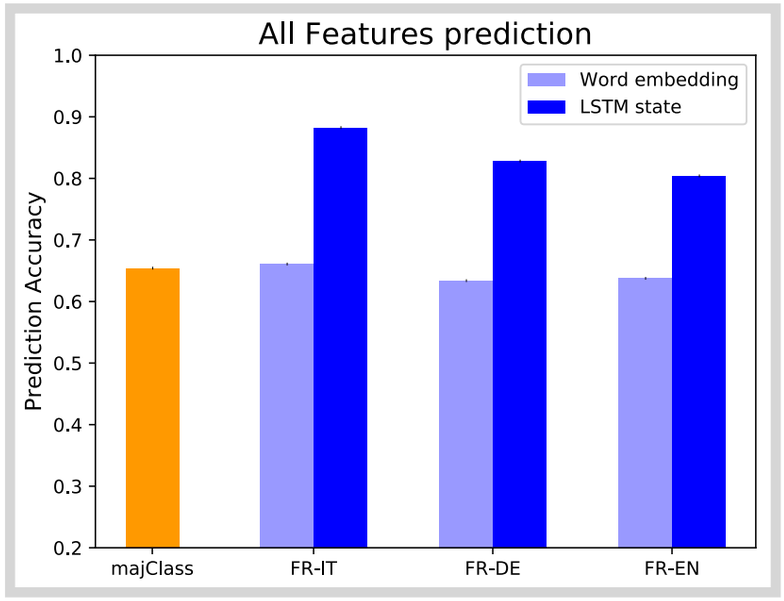
Results (1)

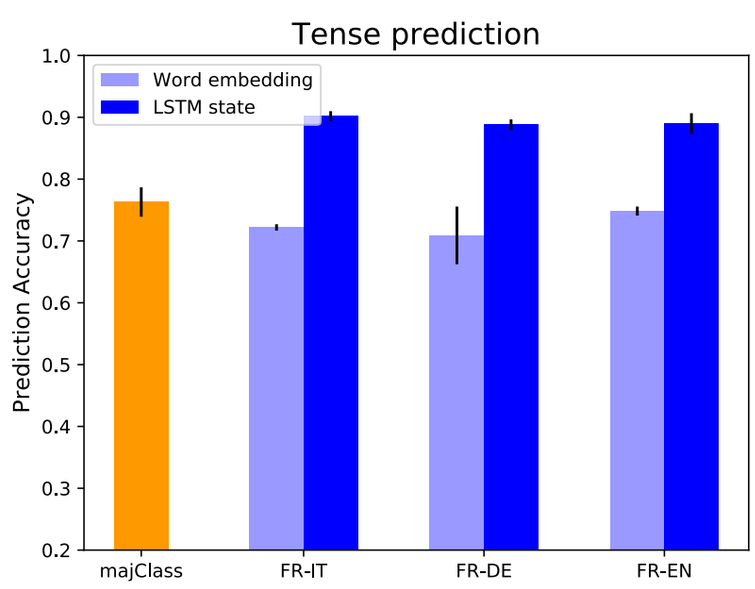
- Are source words' morphological features encoded at all?
- Are some features better captured than others?
- Is morphology captured as a word type property or only in context?

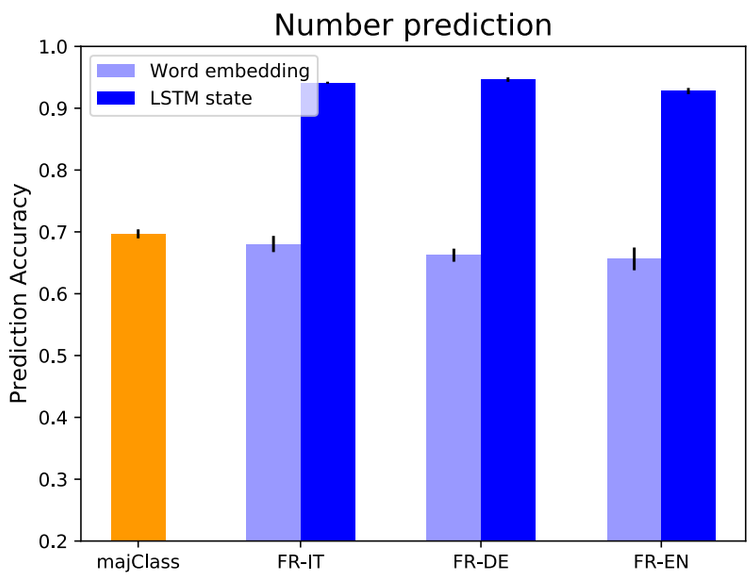


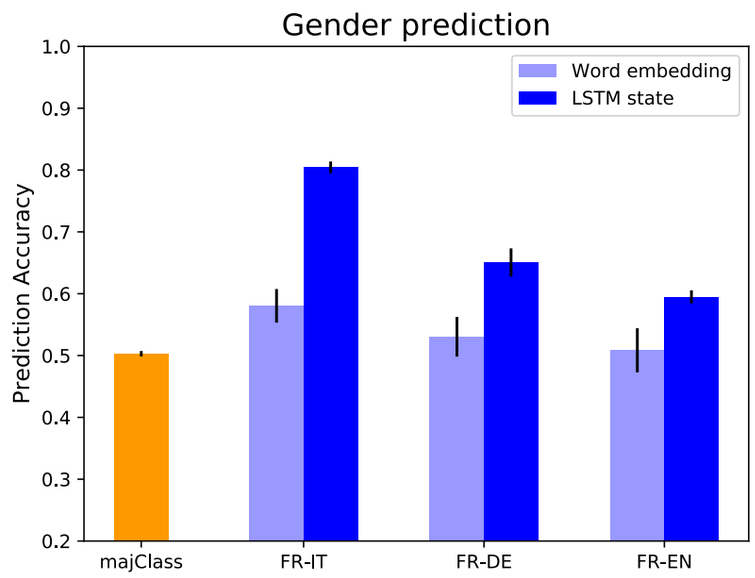
Results (2)

- What's the impact of the target language?
- Does that vary among different morphological features?







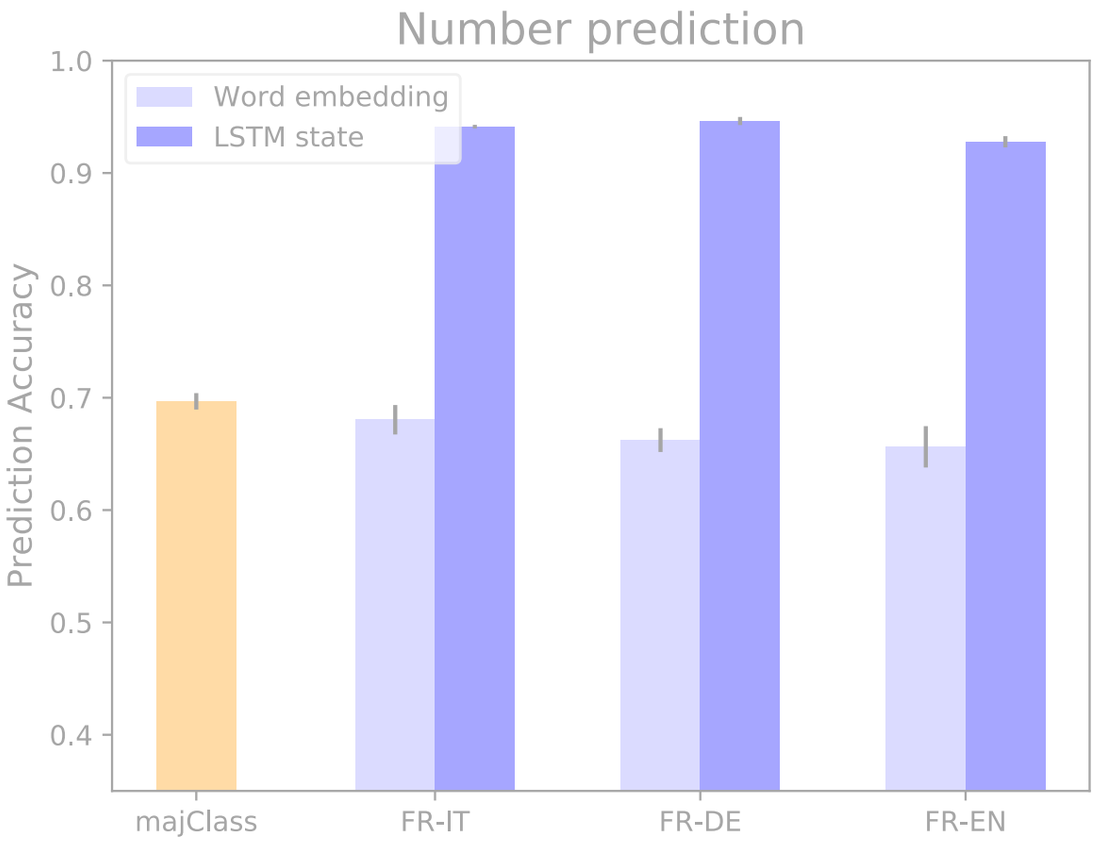


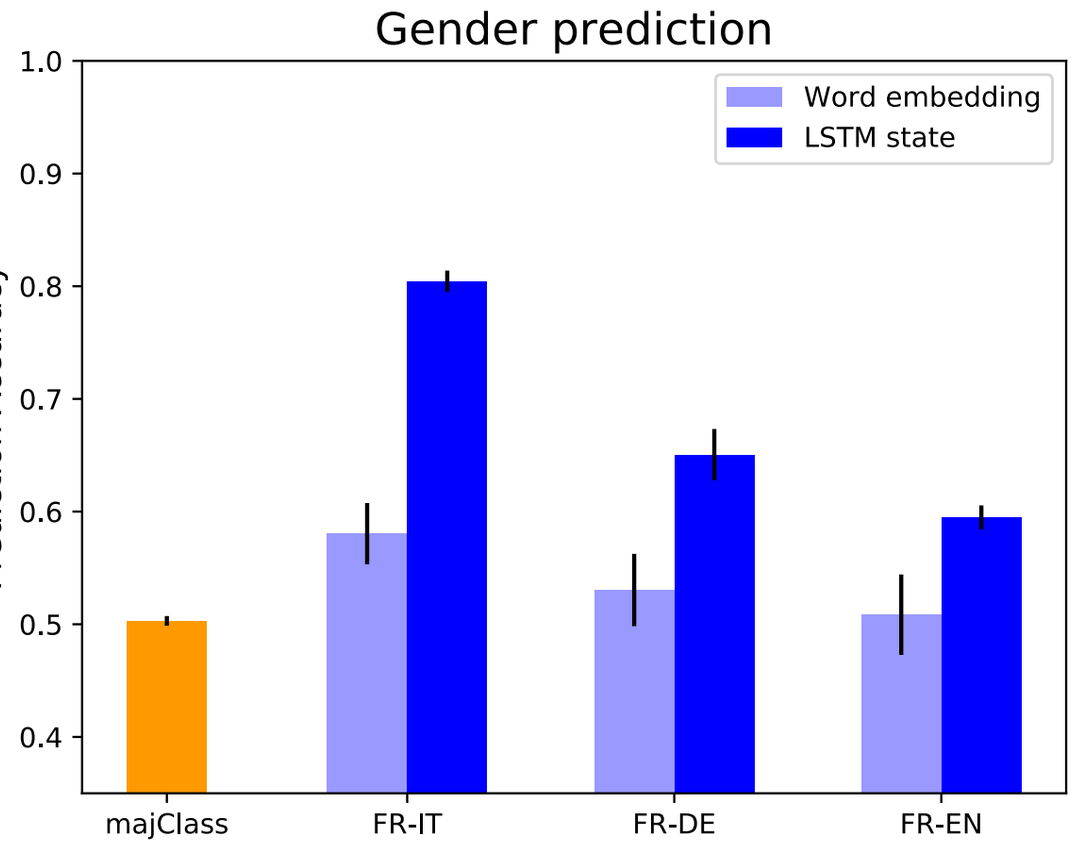
The curious case of Gender

- An intrinsic property of nouns. Triggers agreement in other word classes (adjectives, articles, etc.)
- Present in French, Italian, German, but not in English
- Noun gender is often arbitrary, i.e. no semantic or syntactic value (*cf.* number and case)

The curious case of Gender (2)

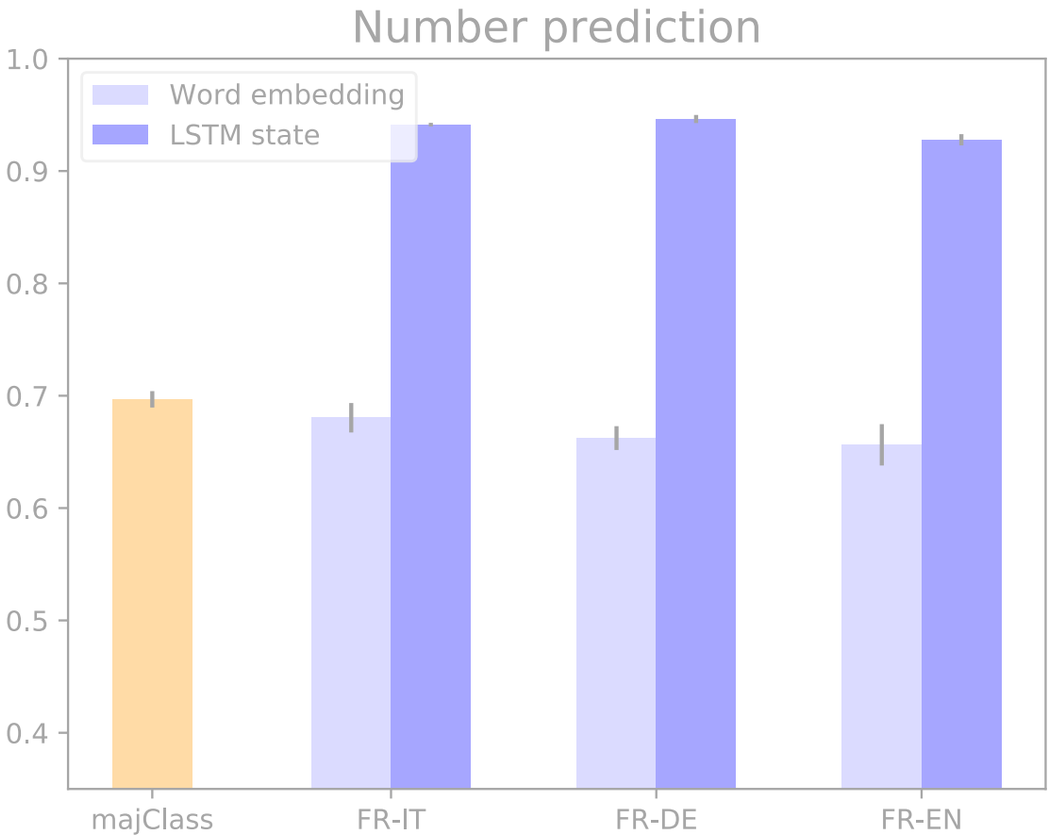
- Explaining factors: language relatedness? gender marking in the target language?

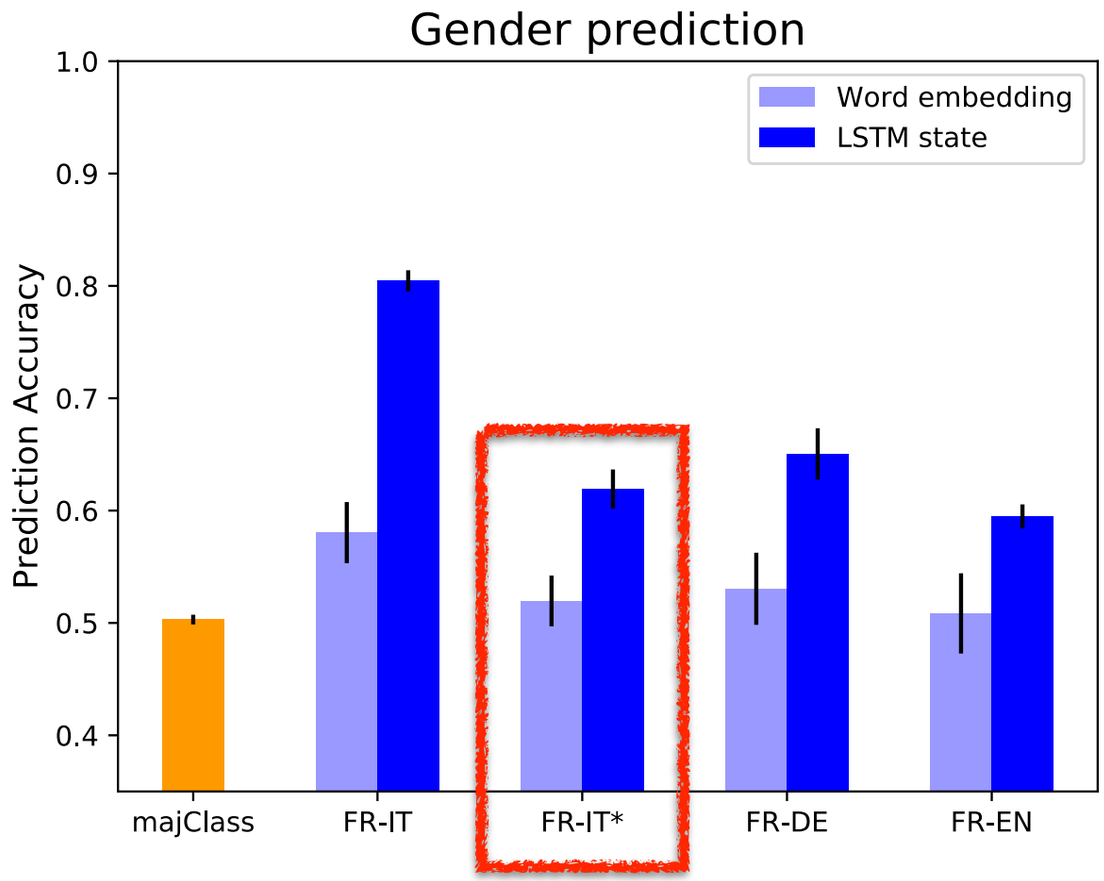




The curious case of Gender (3)

- Explaining factors: language relatedness? gender marking in the target language?
- Experiment with FR-IT*: modified Italian without any gender marking
- Answer: Mostly target gender marking, but not only





Summary

- Nominal and verbal morphology captured to a large extent by NMT encoder, but only in context, not as a word type property

- Semantic features (number, tense) encoded much better than purely grammatical features (gender)

- Gender encoding dramatically affected by target language:

  - explaining factors: (1st) target-side marking, (2nd) language relatedness

  - gender learnt to some extent even without any target-side marking (!)

The field needs …

Interpretable models

The field needs more interpretable models:

- to deliver reliable technology
- to detect limitations and address them
- for scientific interest (does neural translation process resemble human translation or not at all?)

Good linguistic hypotheses

The field *also* needs to ask the right questions:

- the advances of NMT force us more than ever to reason about the object of our study: languages!
- less quantitative, more qualitative evaluation: an age shift?
- much to be done in order to generalize current findings to different phenomena and different types of morphology

Thanks for your attention

Join me in Leiden!
I am looking for a talented PhD student.
Come talk to me if you're interested, or spread the word.