University of Zurich[UZH]

Institute of Computational Linguistics

Document-level context in deep recurrent neural networks

Kolloquium Talk 2017

Mathias Müller

On the menu today

- Establish that document-level context matters for neural machine translation (NMT)
- How to evaluate document-level improvements
- Proposed architecture to integrate arbitrary contexts (multi-context conditional GRU)

- Also on the menu today:

(http://www.mensa.uzh.ch/de/menueplaene/mensa-uzh-binzmuehle/dienstag.html)

Establishing that document-level context matters in NMT

Context matters (rather serious illustration)

Context matters (fabricated example)

Source

The sun is shining. It is bright.

Target

Die Sonne scheint. ____ ist hell.

Context matters (actual WMT examples)

Source

This organism has dual capability. It can grow with either phosphorous or arsenic.

Target

Dieser Organismus hat zwei Möglichkeiten. Er benötigt zum Wachsen entweder Phosphor oder Arsen.

(example taken from `newstest2011.{de,en}`)

Context matters (actual WMT examples)

Sentence-level NMT solves the following task:

Source

This organism has dual capability. It can grow with either phosphorous or arsenic.

Target

Dieser Organismus hat zwei Möglichkeiten. _____ benötigt zum Wachsen entweder Phosphor oder Arsen.

Context matters (actual WMT examples)

Source

However, the European Central Bank (ECB) took an interest in it in a report on virtual currencies published in October. It describes bitcoin as "the most successful virtual currency," […].

Target

Dennoch hat die Europäische Zentralbank (EZB) in einem im Oktober veröffentlichten Bericht über virtuelle Währungen Interesse hierfür gezeigt. Sie beschreibt Bitcoin als "die virtuelle Währung mit dem größten Erfolg" […].

(example taken from `newstest2013.{de,en}`)

Context matters (actual WMT examples)

Source

However, the European Central Bank (ECB) took an interest in it in a report on virtual currencies published in October. It describes bitcoin as "the most successful virtual currency," […].

Target

Dennoch hat die Europäische Zentralbank (EZB) in einem im Oktober veröffentlichten Bericht über virtuelle Währungen Interesse hierfür gezeigt. ____ beschreibt Bitcoin als "die virtuelle Währung mit dem größten Erfolg" […].

Context matters (actual WMT examples)

Google

Translate

English | French | German | Detect language

However, the European Central Bank (ECB) took an interest in it in a report on virtual currencies published in October. It describes bitcoin as "the most successful virtual currency".

183/5000

French | German | English | Translate

Die Europäische Zentralbank (EZB) interessierte sich dagegen in einem Bericht über virtuelle Währungen, der im Oktober veröffentlicht wurde. Es beschreibt Bitcoin als "die erfolgreichste virtuelle Währung".

Suggest an edit

Do we treat NMT models fairly?

Source

It describes bitcoin as "the most successful virtual currency”.

Target

Es beschreibt den Bitcoin als "die erfolgreichste virtuelle Währung".

Establishing that document-level context matters in NMT

How to evaluate document-level improvements

How to evaluate automatically?

- Metrics like BLEU too coarse-grained
- Also, impossible to focus evaluation on specific linguistic phenomena

Solutions:

- Use specialized metrics (Miculicich Werlen and Popescu-Belis, 2017)
- Design “challenge sets”, for contrastive evaluation

Challenge set evaluation

- Idea: take advantage of the fact that NMT systems are *conditional language models*
- Contrastive evaluation by model scoring:

Source
Despite the fact that it is a part of China, Hong Kong determines its currency policy separately.

Target
Hongkong bestimmt, obwohl **es** zu China gehört, seine Währungspolitik selbst.

Contrastive
Hongkong bestimmt, obwohl **er** zu China gehört, seine Währungspolitik selbst.

(example taken from `newstest2009`)

Challenge set evaluation

Previous experience with challenge sets:

- hand-selected, manually annotated examples to test pronoun translation (Guillou and Hardmeier, 2016)
- first application to NMT: LingEval97 (Sennrich, 2017)
- extension to words with several senses: ContraWSD (Rios et al., 2017)

And, very recently:

- name “challenge set” due to Isabelle et al. (2017)
- handcrafted set with ambiguous pronouns: Bawden (in preparation)

Contra Pronoun Challenge Set


```
{
    "antecedent head": "Gladbacher",
    "antecedent head gender": "Masc",
    "antecedent head lemma": "Gladbacher",
    "antecedent head pos": "NN",
    "antecedent phrase": "Der Gladbacher",
    "document id": "newstest2009-ref.de.19",
    "errors": [
        {
            "contrastive": "Sie wusste schlichtweg, dass er Recht hatte.",
            "gender of replacement": "Fem",
            "replacement": "Sie",
            "type": "pronominal coreference"
        },
        {
            "contrastive": "Es wusste schlichtweg, dass er Recht hatte.",
            "gender of replacement": "Neut",
            "replacement": "Es",
            "type": "pronominal coreference"
        }
    ],
    "intrasentential": false,
    "origin": "newstest2009-ende",
    "reference": "Er wusste schlichtweg, dass er Recht hatte.",
    "reference pronoun": "Er",
    "segment id": 7,
    "source": "He simply knew that he was right.",
    "source pronoun": "He"
},
```

Establishing that document-level context matters in NMT

How to evaluate document-level improvements

Proposed architecture to integrate arbitrary contexts (multi-context conditional GRU)

Integrating document-level context

Into existing architectures:

- Nematus, an extremely successful tool (Sennrich et al., 2017)
- encoder-decoder network with soft attention (Bahdanau et al., 2014)
- encoder and decoder are recurrent neural networks (RNNs)

Rule out simple solutions:

- concatenate sentences problematic because of sequence length (Koehn and Knowles, 2017)

What are other groups doing?

Known NMT solutions that have intersentential context:

- gated auxiliary context or ”warm start” decoder initialization with a document summary (Wang et al., 2017)
- additional encoder and attention network for previous source sentence (Jean et al., 2017)
- Concatenate previous source sentence, mark with a prefix (Tiedemann and Scherrer, 2017)
- both source and target context (Miculicich Werlen et al., under review)

Actual Implementation

Building on previous work,

- extension of conditional gated recurrent unit (cGRU) RNN that Nematus uses as decoder
- allow for arbitrary *past* (obviously) context sizes, both source and target side
- 1 additional encoder for each context, 1 additional GRU unit with attention during *deep transition*

Recurrent neural networks refresher

$$\mathbf{h}_t = \sigma_h(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h \mathbf{h}_{t-1} + \mathbf{b}_h)$$

$$\mathbf{y}_t = \sigma_y(\mathbf{W}_y \mathbf{h}_t + \mathbf{b}_y)$$

RNN variant: gated recurrent unit (GRU)

$$\mathbf{h}_t = (1 - \mathbf{z}_t)\mathbf{h}_{t-1} + \mathbf{t}_t\tilde{\mathbf{h}}_t$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1})$$

$$\tilde{\mathbf{h}}_t = tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1}))$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1})$$

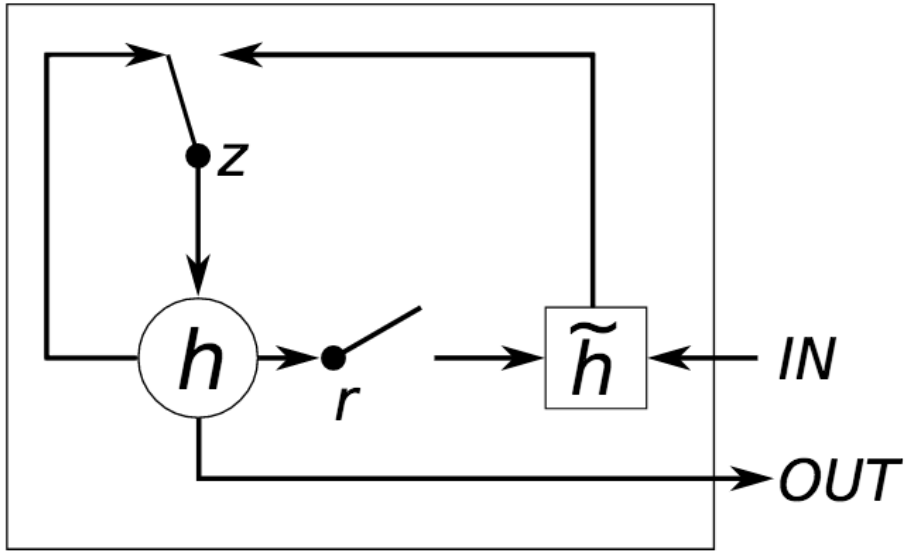

Figure taken from Chung et al. (2014)

Notion of *depth* in RNN networks

- generally three types of depth (Pascanu et al., 2013):

stacked layers	(each layer individually recurrent)
deep transition	(units not individually recurrent)
deep output	(units not individually recurrent)

- in Nematus, the decoder is implemented as a cGRU with deep transition and deep output
- crucially: attention over source sentence vectors **C** is a deep transition step

Conditional gated recurrent unit (cGRU)

$$\mathbf{h}_t = cGRU_{att}(\mathbf{h}_{t-1}, \mathbf{y}_{t-1}, \mathbf{C})$$

$$\mathbf{h}_t^1 = GRU_1(y_{t-1}, \mathbf{h}_{t-1})$$

$$\mathbf{c}_t = ATT(\mathbf{C}, \mathbf{h}_t^1)$$

$$\mathbf{h}_t = GRU_2(\mathbf{h}_t^1, \mathbf{c}_t)$$

Detailed formulas: https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf

Extension of cGRU for n contexts

$$\mathbf{h}_t = cGRU_{att}(\mathbf{h}_{t-1}, y_{t-1}, \mathbf{C})$$

$$\mathbf{h}_t = 2cGRU_{att}(\mathbf{h}_{t-1}, y_{t-1}, \mathbf{C}_{main}, \mathbf{C}_{aux})$$

$$\mathbf{h}_t = ncGRU_{att}(\mathbf{h}_{t-1}, y_{t-1}, \mathbf{C}_1, ..., \mathbf{C}_n)$$

Detailed formulas: https://github.com/bricksdont/ncgru/blob/master/ct.pdf

Outlook

Experiments with this new architecture until end of the year:

- small source context vs. equally deep baseline
- target context seems to be useful (Bawden, in preparation)
- challenge set evaluation, focus on pronouns
- use attention as an inspection tool (Kuncoro et al., 2016; Rikters et al., 2017)

Then, look for more general solution, maybe outside of Nematus

- investigate other kinds of networks: fully convolutional (Gehring et al., 2017) or self-attention ("transformer") models (Vaswani et al., 2017), both with positional embeddings

Thanks!

Code currently here:
https://gitlab.cl.uzh.ch/mt/nematus-context2