Temporal Machine Translation

Atila Martens

January 16, 2024

Institut für Computerlinguistik

Introduction

Language changes

- human language evolves
- historical documents are hard to comprehend
    - limits accessibility to scholars specialized in the time period

Examples

Wycliffe:
Blessid ben merciful men, for thei schulen gete merci.

Tyndale:
Blessed are the mercifull: for they shall obteyne mercy.

King James:
Blessed are the mercifull: for they shall obtaine mercie.

Examples

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date:
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimm'd;
And every fair from fair sometime declines,
By chance or nature's changing course untrimm'd;
But thy eternal summer shall not fade
Nor lose possession of that fair thou ow'st;
Nor shall Death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st;
So long as men can breathe or eyes can see,
So long lives this, and this gives life to thee.

Shall I compare you to a summer day?
You're lovelier and milder.
Rough winds shake the pretty buds of May,
and summer doesn't last nearly long enough.
Sometimes the sun shines too hot,
and often its golden face is darkened by clouds.
And everything beautiful stops being beautiful,
either by accident or simply in the course of nature.
But your eternal summer will never fade,
nor will you lose possession of your beauty,
nor shall death brag that you are wandering in the underworld,
once you're captured in my eternal verses.
As long as men are alive and have eyes with which to see,
this poem will live and keep you alive.

Difficulties

Why

- lack of:
  - spelling convention
  - ortography changes depending on author & time

Spelling normalization

- one proposed solution
- adapt documents spelling to modern standards
- orthography consistency & better readability

Spelling normalization

	ENHG	NORM	MOD	Translation
(i)	alle	alle	✓	all
	schult	Schuld	^{2}Schulden	debts
	die	die	✓	that
	die	die	✓	the
	fraw	Frau	✓	woman
	und	und	✓	and
	ire	ihre	✓	her
	kyndt	Kind	^{2}Kinder	children
	schuldig	schuldig	✓	owing
	sint	sind	✓	are

Previous work

- spelling normalization & modernization using a set of rules
- SMT - statistical machine translation

This work

- NMT - neural machine translation & SMT
- creation of two small corpora using backtranslation

Approaches

Goal

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y} \mid \mathbf{x})$$

SMT & NMT

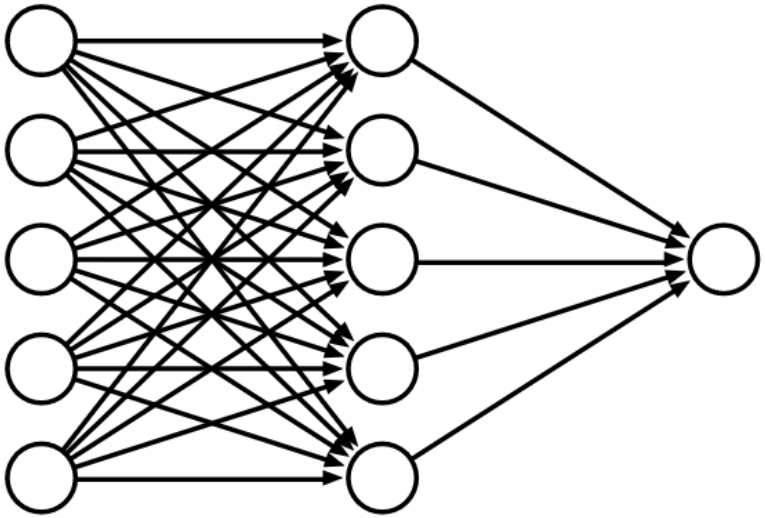
- for years the best method to use were phrase-based models
- shifted slowly to neural models
  - frequently relies on RNN endoder-decoder networks
  - this paper uses an LSTM encoder-decoder network with an attention layer in between

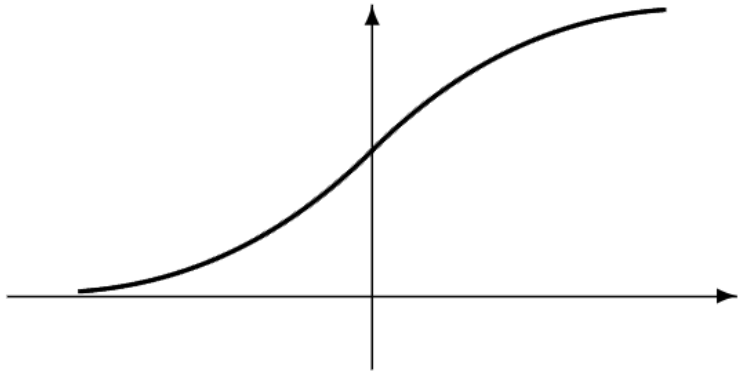
FFNN

$$\mathbf{h} = f(\mathbf{x}\mathbf{W}^{(1)}),$$

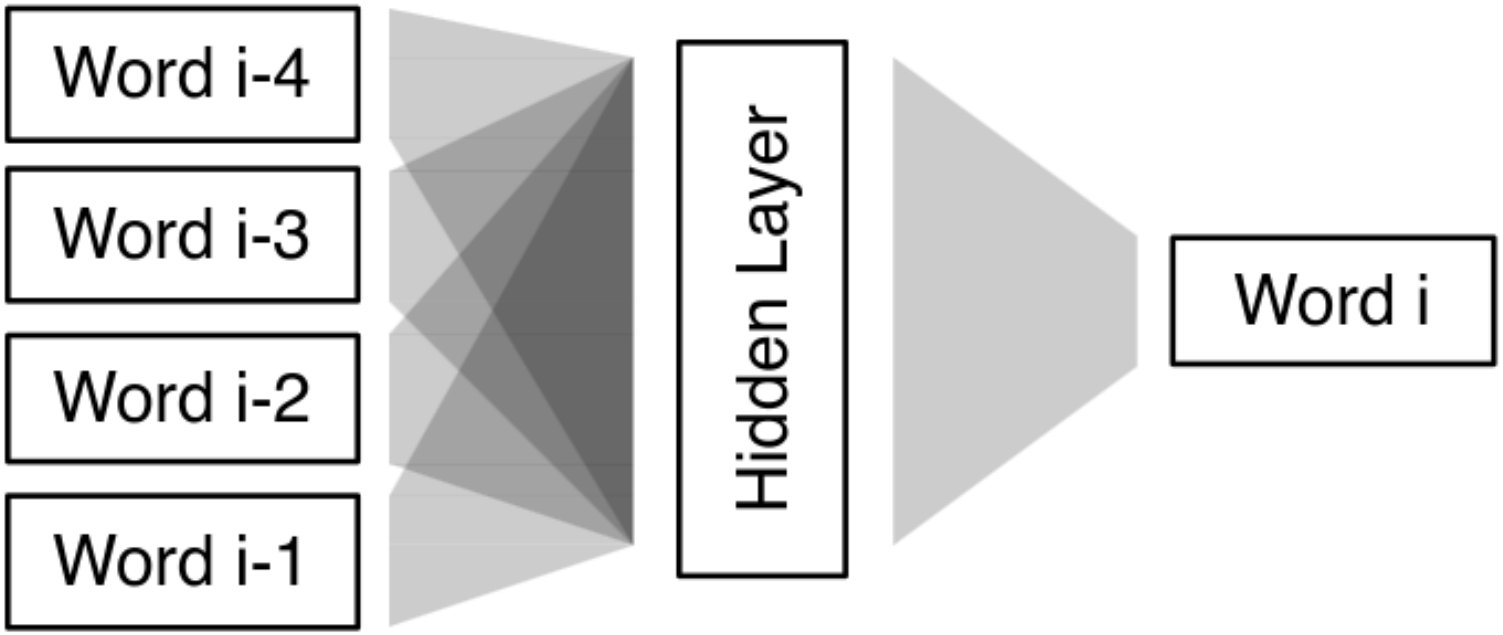
$$\mathbf{y} = g(\mathbf{h}\mathbf{W}^{(2)})$$

Logistic function
$\text{sigmoid}(x) = \sigma(x) = \frac{1}{1+e^{-x}}$
output ranges from 0 to $+1$

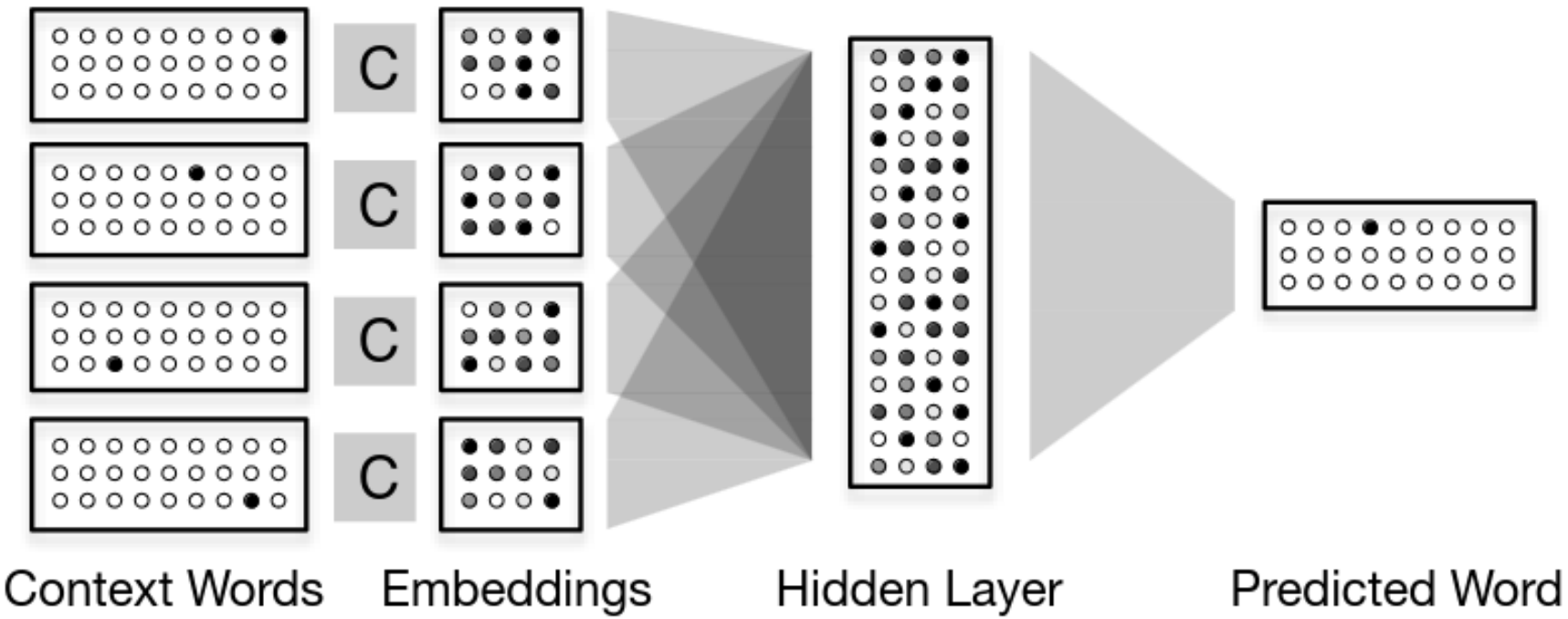




FFNN

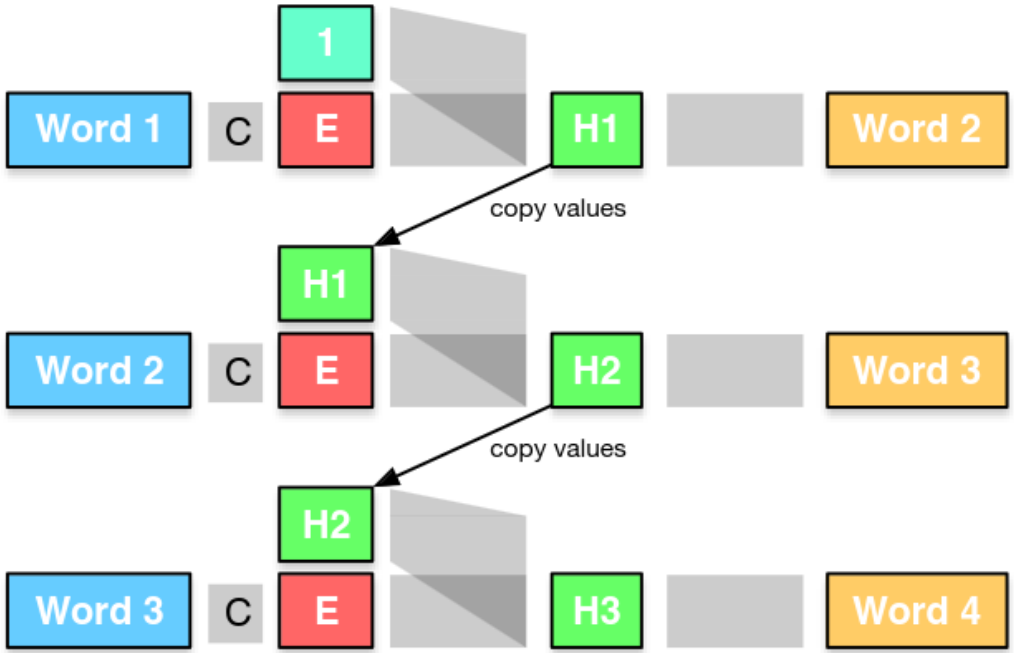


FFNN


C
C
C
C
Context Words
Embeddings
Hidden Layer
Predicted Word

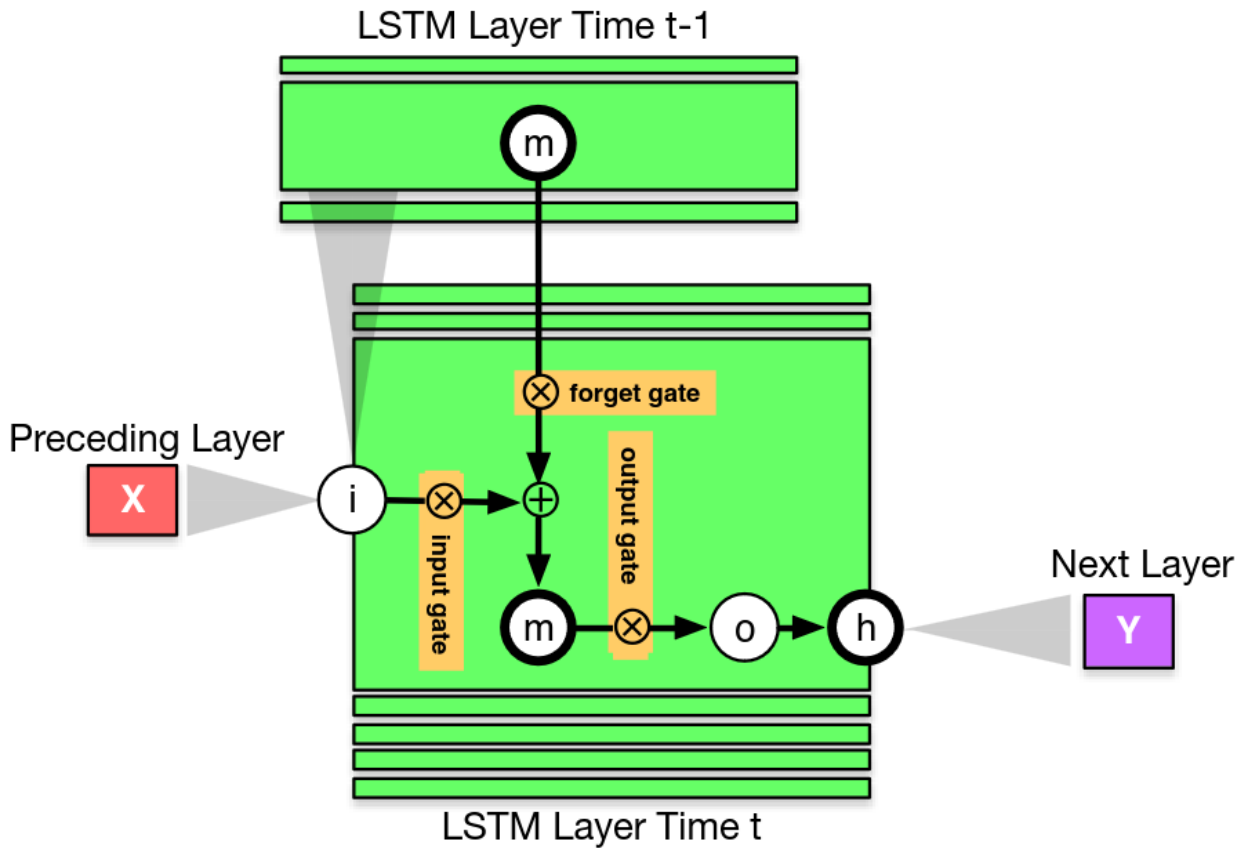
RNN

$$\begin{aligned}\mathbf{h}_t &= f(\mathbf{x}_t, \mathbf{h}_{t-1}) \\ &= \sigma(\mathbf{x}_t \mathbf{W}^{(\mathbf{x1})} + \mathbf{h}_{t-1} \mathbf{W}^{(\mathbf{h1})}), \\ \mathbf{y}_t &= \text{softmax}(\mathbf{h}_t \mathbf{W}^{(\mathbf{h2})}).\end{aligned}$$

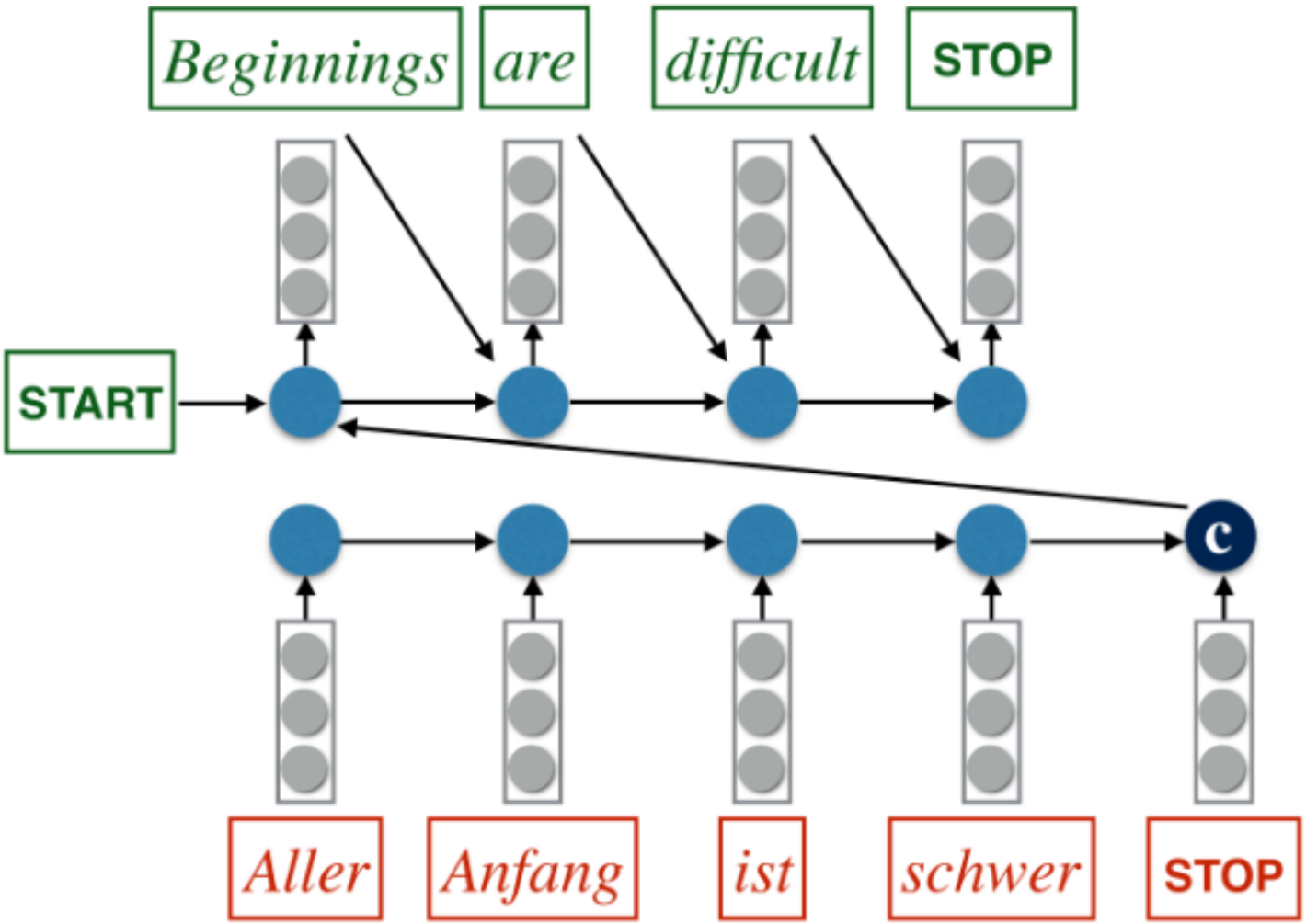


LSTM

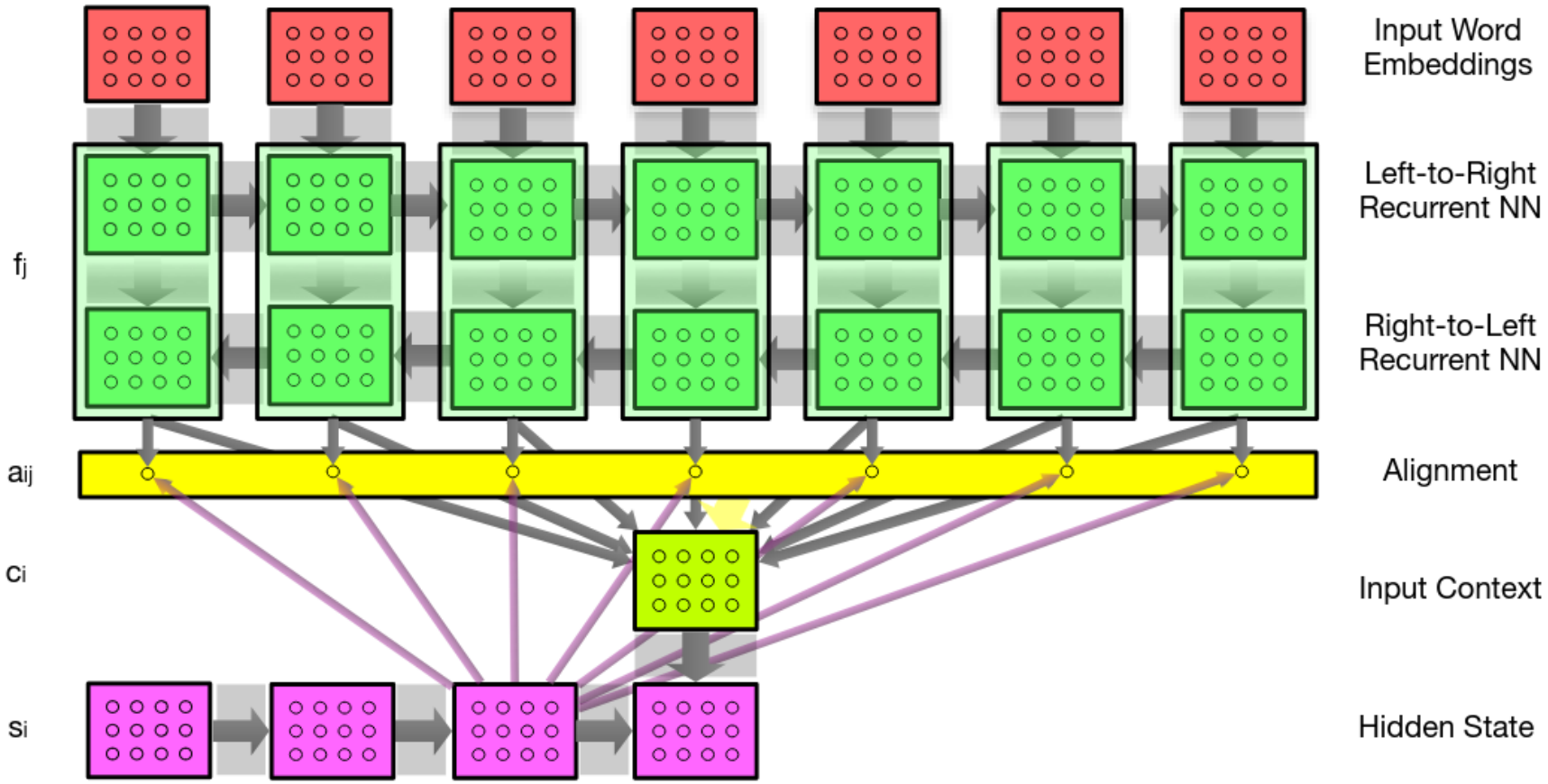
$$
\begin{aligned}
\mathbf{h}_t &= f(\mathbf{x}_t, \mathbf{h}_{t-1}) \\
&= \tanh(\mathbf{m}_t \otimes \mathbf{o}), \\
\mathbf{m}_t &= \mathbf{m}_{t-1} \otimes \mathbf{f} + \mathbf{g} \otimes \mathbf{i}, \\
\mathbf{i} &= \sigma(\mathbf{x}_t \mathbf{W}^{(\mathbf{xi})} + \mathbf{h}_{t-1} \mathbf{W}^{(\mathbf{hi})}), \\
\mathbf{f} &= \sigma(\mathbf{x}_t \mathbf{W}^{(\mathbf{xf})} + \mathbf{h}_{t-1} \mathbf{W}^{(\mathbf{hf})}), \\
\mathbf{o} &= \sigma(\mathbf{x}_t \mathbf{W}^{(\mathbf{xo})} + \mathbf{h}_{t-1} \mathbf{W}^{(\mathbf{ho})}), \\
\mathbf{g} &= \tanh(\mathbf{x}_t \mathbf{W}^{(\mathbf{xg})} + \mathbf{h}_{t-1} \mathbf{W}^{(\mathbf{hg})})
\end{aligned}
$$



RNN encoder-decoder



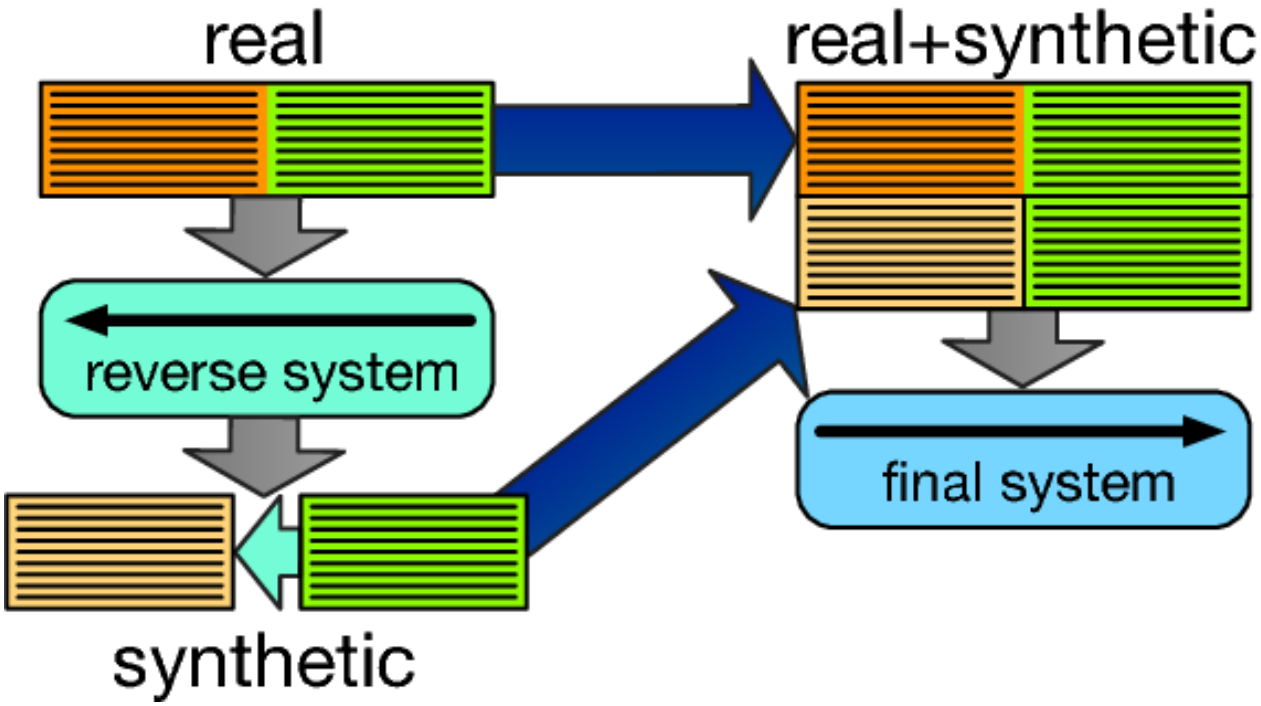
LSTM encoder-decoder



Backtranslation

- creating synthetic text from monolingual data
- ad-hoc SMT system trained on the corpus' training partition translated data
- had to replicate training data several times to match size of synthetic data

Backtranslation



Dutch bible

- different versions of the bible in dutch (1637, 1657, 1888, 2010)
- used the 1637 version as the original document
- and the 1888 version as the modern dutch version
- for the synthetic corpus: collected all 19^{th} century dutch books from an online library

El Quijote

- 17^{th} century spanish novel as original & 21^{st} modernized version
- 17^{th} century is faithful to the original manuscript (only a couple of words per line)
  - replaced line breaks with spaces
  - removed empty lines
  - added line breaks at relevant punctuation
  - same for 21^{st} century version for consistency
- aligned both documents
- for synthetic data: monolingual data from spanish literature

El Conde Lucanor

- original 14^{th} century novel & 21^{st} century modernized version
- same pre-processing as for El Quijote
- was too small, so only used as test set

Corpora statistics

		Dutch Bible	El Quijote	El Conde Lucanor
Train	$\|S\|$	35.2K	10K	-
	$\|T\|$	870.4/862.4K	283.3/283.2K	-
	$\|V\|$	53.8/42.8K	31.7/31.3K	-
Development	$\|S\|$	2000	2000	-
	$\|T\|$	56.4/54.8K	53.2/53.2K	-
	$\|V\|$	9.1/7.8K	10.7/10.6K	-
Test	$\|S\|$	5000	2000	2252
	$\|T\|$	145.8/140.8K	41.8/42.0K	62.0/56.7K
	$\|V\|$	10.5/9.0K	8.9/9.0K	7.4/8.6K
Monolingual	$\|S\|$	4.1M	567.0K	-
	$\|T\|$	88.3M	9.5M	-
	$\|V\|$	2.0M	470.4K	-

|S| = Sentences, |T| = Tokens, |V| = Vocabulary

Results

Metrics

- BLEU: **B**i**L**ingual **E**valuation **U**nderstudy
  - geometric average of modified n-gram precision
  - brevity factor that penalizes short sentences
- TER: **T**ranslation **E**rror **R**ate
  - computes number of word edit operations
  - normalized by number of words in final translation

Comparison

System	Dutch Bible		El Quijote		El Conde Lucanor	
	BLEU	TER	BLEU	TER	BLEU	TER
Baseline	13.5 ± 0.3	57.0 ± 0.3	36.5 ± 0.8	43.3 ± 1.1	5.8 ± 0.3	89.6 ± 1.0
Baseline_2	50.8 ± 0.4	26.5 ± 0.3	-	-	-	-
SMT	$\mathbf{80.1 \pm 0.5}$	$\mathbf{9.9 \pm 0.3}$	$\mathbf{58.9 \pm 1.0}$	$\mathbf{29.4 \pm 1.2}$	$\mathbf{8.4 \pm 0.3}$	$\mathbf{83.8 \pm 1.0}$
NMT	38.0 ± 0.6	51.7 ± 2.2	37.4 ± 1.2	51.5 ± 2.0	2.7 ± 0.2	99.5 ± 2.0
$\text{NMT}_{\text{Synthetic}}$	17.4 ± 0.5	65.6 ± 1.7	45.2 ± 1.3	50.6 ± 3.5	3.1 ± 0.2	165.1 ± 8.2

Example sentences

El Quijote

Original:	Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lança atrauessada por la boca, con la fiereça que suele pintarse.
Modernized:	Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.
SMT:	Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.
NMT:	Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;
NMT$_{\text{Synthetic}}$:	Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fierded que suele pintarse.

Example sentences

El Conde Lucanor

Original:	-Señor conde Lucanor -dixo Patronio-, vien entiendo que el mío consejo non vos faze grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, fazerlo he luego.
Modernized:	-Señor Conde Lucanor -dijo Patronio-, bien sé que mi consejo no os hace mucha falta, pero, como confiáis en mí,
SMT:	–– Señor conde Lucanor -dijo Patroniorosa, vien entiendo que el mío consejo non vos face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos aconseje en ello, ferlo he luego .
NMT:	Señor conde Olcanor dijo dijo Pacasos –dijo en entiendo que el mío consejo non os fazo felimengua y vuestra merced es que vos diga lo que en esto entiendo.
NMT$_{\text{Synthetic}}$:	-Señor conde Lucanor -dijo Patronio, vien entiendo que el mío consejo non es face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, también yo he dicho.

Thoughts

Devastating conclusion

I'm not the biggest fan of this paper:

- not convinced the presented problem is as dire as they make it seem
- furthermore it is almost certain that parts of the documents intentions are lost, so why bother
- they used a metric that penalizes good output
- they used a baseline that is not really a baseline
    - and still managed to get worse results
- a lot of spelling mistakes in the paper

Discussion

1. Why did they not build a bigger corpus with multiple books in it?
2. Do you think their baseline makes any sense?
3. Do you know of a better metric they could have used?