Natural Language Processing and Deep Learning

Alexander Rush
(with Yoon Kim, Sam Wiseman, Yacine Jernite,
Jason Weston, Sumit Chopra, David Sontag, Stuart Shieber)

Smoothness Image/Language

It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts. -Sherlock Holmes, A Scandal in Bohemia

Smoothness Image/Language

It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts. -Sherlock Holmes, A Scandal in Bohemia

Repair Image / Language (Chatterjee et al., 2009)

It is a capital mistake to theorize before one has ______ . . .

Repair Image / Language (Chatterjee et al., 2009)

108 938 285 28 184 29 593 219 58 772 ______ . . .

Language Modeling

Learn distribution from data:

$$p(w_{t+1}|w_1, \ldots, w_t)$$

Language Modeling

Learn distribution from data:

$$p(w_{t+1}|w_1, \ldots, w_t)$$

- Speech Recognition
- Machine Translation
- Summarization
- Dialogue
- Soft Keyboards
- Word Correction
- Text Simplification
- ...

Language Modeling Recipe (pre-2010)

Goal: Estimate n-gram model (Markov assumption)

$$p(w_{t+1}|w_1, \dots, w_t) \approx p(w_{t+1}|w_{t-n+1}, \dots w_t)$$

Ingredients:

- 1 Corpus (e.g. the entire web)

Steps:

- (1) Collect words, (2) Count up n-grams, (3) Divide*

$$\begin{aligned} p(w_{t+1}|w_{t-n+1}, \dots w_t) &= \frac{\#(w_{t-n+1}, \dots, w_{t+1})}{\#(w_{t-n+1}, \dots, w_t)} \\ &= \frac{\#(\text{theorize before one has data})}{\#(\text{theorize before one has})} \end{aligned}$$

How Good Is a Language Model?

Perplexity:

$$\exp(-\sum_{t=1}^{T} \frac{1}{T} \log p(w_{t+1} | w_1, \ldots, w_t))$$

- corresponds to size of equally predictive uniform distribution

On Wall Street Journal (PTB):

- Vocabulary $|\mathcal{V}| = 10{,}000$ word types
- Words $T \approx 1$ million

Language Model	Perplexity
Uniform	10000
KN / 5-gram	141

Deep Learning for Language Modeling
(Bengio et al., 2003), (Mikolov et al., 2010)

Recurrent neural network (RNN) models estimate (non-Markovian):

It is a capital mistake to theorize before one has data

$$p(w_{t+1}|w_1, \ldots, w_t)$$

Long-Short Term Memory (LSTM)(Hochreiter and Schmidhuber, 1997) RNN language models

Language Model	Perplexity
Uniform	10000
KN / 5-gram	141
LSTM (Zaremba et al., 2014)	78(!)

Deep Learning for Language Modeling

(Bengio et al., 2003), (Mikolov et al., 2010)

Recurrent neural network (RNN) models estimate (non-Markovian):

It is a capital mistake to theorize before one has data

$$p(w_{t+1}|w_1, \ldots, w_t)$$

Long-Short Term Memory (LSTM)(Hochreiter and Schmidhuber, 1997) RNN language models

Language Model	Perplexity
Uniform	10000
KN / 5-gram	141
LSTM (Zaremba et al., 2014)	78(!)

Idea 1: From Discrete Elements to Embeddings

Learn input embeddings (vectors) for each words in vocab.

$$\mathbf{U} \in \mathbb{R}^{|\mathcal{V}| \times D}, D \approx 256$$

Example:

w	$\mathbf{U}_w$				
theorize	[0.2,	-0.2,	-0.1,	0.4,	-0.2, ...]
before	[0.0,	0.3,	-0.4,	-0.3,	0.0, ...]
one	[0.1,	-0.2,	-0.1,	-0.0,	-0.2, ...]
has	[0.5,	-0.1,	0.1,	0.3,	0.3, ...]
...					

Idea 2: From Embeddings to Representations

Combine input vectors into an hidden representation of context.

$$\mathbf{h}_0 = \mathbf{0}$$

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{U}_{w_t}) \text{ for all } \ t > 0$$

Example:

$$\begin{tikzcd}
\boxed{\cdots} \arrow[r] & \boxed{\mathbf{h}_{t-2}} \arrow[r] & \boxed{\mathbf{h}_{t-1}} \arrow[r] & \boxed{\mathbf{h}_t} \\
 & \boxed{\mathbf{U}_{\text{before}}} \arrow[u] & \boxed{\mathbf{U}_{\text{one}}} \arrow[u] & \boxed{\mathbf{U}_{\text{has}}} \arrow[u]
\end{tikzcd}$$

Idea 3: From Representation to Output Embedding

Learn output embeddings (and bias) for each word in vocab:

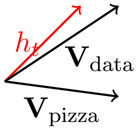
$$\mathbf{V} \in \mathbb{R}^{|\mathcal{V}| \times D}, \mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$$

Score of word w is dot-product with hidden representation .

$$s(w) = \mathbf{V}_w^\top \mathbf{h}_t + \mathbf{b}_w$$

Example:

$$\mathbf{V}_{\text{pizza}}^\top \mathbf{h}_t \leq \mathbf{V}_{\text{data}}^\top \mathbf{h}_t$$

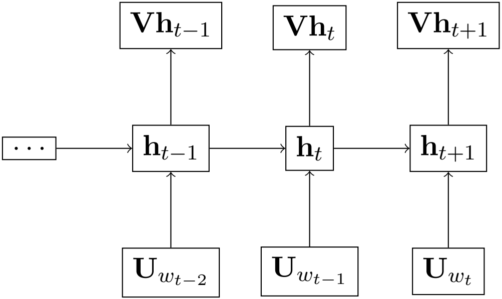


Putting it together

- Apply *soft-max* to convert to probability distribution

$$p(w_{t+1}|w_1,\ldots,w_t) = \frac{\exp(s(w_{t+1}))}{\sum_{w'\in\mathcal{V}}\exp(s(w'))}$$

- Whole model trained together on a large corpus
- Backpropagation with stochastic gradient descent.



Caveats

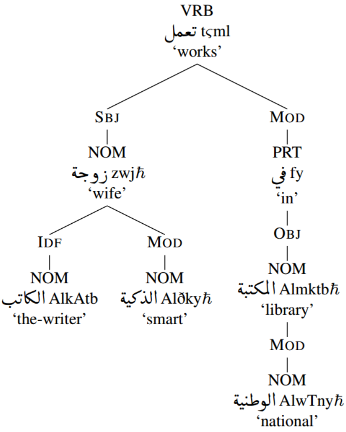
- Combination function for LSTM ($f(\mathbf{h}_{t-1}, w_t)$) is quite complex.

$$
\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}^i \mathbf{x}_t + \mathbf{U}^i \mathbf{h}_{t-1} + \mathbf{b}^i) \\
\mathbf{f}_t &= \sigma(\mathbf{W}^f \mathbf{x}_t + \mathbf{U}^f \mathbf{h}_{t-1} + \mathbf{b}^f) \\
\mathbf{o}_t &= \sigma(\mathbf{W}^o \mathbf{x}_t + \mathbf{U}^o \mathbf{h}_{t-1} + \mathbf{b}^o) \\
\mathbf{g}_t &= \tanh(\mathbf{W}^g \mathbf{x}_t + \mathbf{U}^g \mathbf{h}_{t-1} + \mathbf{b}^g) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t)
\end{aligned}
$$

- Model is non-linear and training objective non-convex.
- Requires hyper-parameter tuning and clever regularization.
- Training is computationally very difficult (use GPUs).

Our Motivation: Trimming the Language Pipeline

تعمل زوجة الكاتب الذكية في المكتبة الوطنية



(Marton et al., 2010)

- Morphological Seg.
- Morphological Tagging
- Part-of-Speech
- Entity Recognition
- Syntactic Parsing
- Role Labeling
- Discourse Analysis

Our Motivation: Structure from Data

- Can this explicit structure can be learned latently from data?
- What architectural elements support our learning linguistic representations?

Projects:

- Character-Aware Language Models [CharCNN]
- Sentence Summarization [Contextual Attention]
- Coreference Resolution [Feature Embeddings]

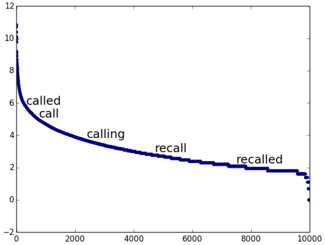
Character-Aware Language Models (Kim et al., 2015)

(1) Character-Aware Language Models

Issue: Embeddings $\mathbf{U}$ different for: “called”, “call”, “calling”, “recalling” and “recalled”.

Goal: Extend recurrent language model to exploit character structure.

- Share properties for “close” words.



- Capture syntactic aspects of morphologically-rich languages.

Past Work

Require preprocessing of morphological segmentation

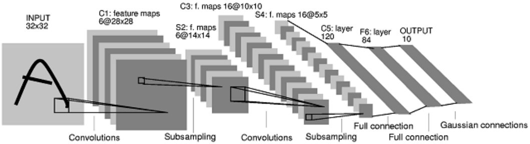
recalling $\Rightarrow$ re - call - ing

- Alexandrescu and Kirchhoff (2006); Bilmes and Kirchhoff (2003): Factored Language models with morphology.
- Luong et al. (2013): LM with Recursive NN over morpheme embeddings
- Botha and Blunsom (2014): LBL with sum over word/morpheme embeddings.

Convolutional Neural Networks (CNN)
(LeCun et al., 1989)

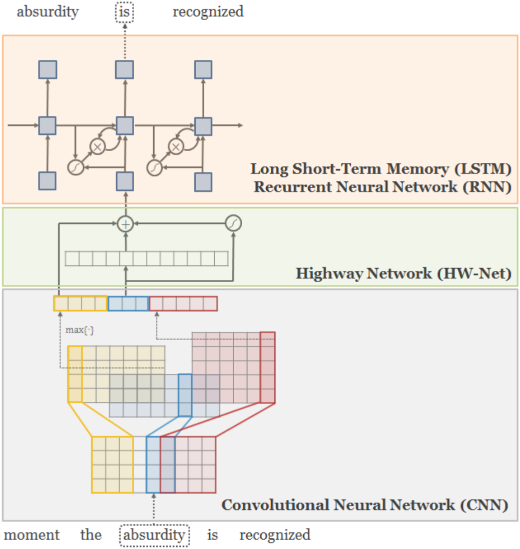
Main Idea: No morphology, use characters directly.

- Central network architecture of deep learning in vision.



- Used for NLP tasks, often over the words. (Collobert et al., 2011; Kalchbrenner et al., 2014; Kim, 2014)

Convolution into Recurrent Model



Character Convolution (CharCNN)

a b s u r d i t y

Character Convolution (CharCNN)

$\mathbf{Q} \in \mathbb{R}^{|\mathcal{C}| \times D}$: Matrix of character embeddings

0.4	-0.8	2.2	0.1	0.5	-0.4	0.4	-0.4	0.1
0.1	1.2	1.5	-0.8	-1.5	0.2	0.1	1.2	0.7
0.2	0.1	-1.2	0.2	-0.2	0.3	0.2	-1.3	-0.1
-0.2	-0.5	0.1	0.2	-0.3	0.3	-0.1	1.0	-0.3

. . . .

a b s u r d i t y

Character Convolution (CharCNN)

$\mathbf{H} \in \mathbb{R}^{D \times w}$: Convolutional filter matrix of width $w = 3$

-0.1	0.5	2.2
0.7	0.9	0.3
-0.2	-0.2	0.7
1.3	-0.1	-1.1

0.4	-0.8	2.2	0.1	0.5	-0.4	0.4	-0.4	0.1
0.1	1.2	1.5	-0.8	-1.5	0.2	0.1	1.2	0.7
0.2	0.1	-1.2	0.2	-0.2	0.3	0.2	-1.3	-0.1
-0.2	-0.5	0.1	0.2	-0.3	0.3	-0.1	1.0	-0.3

. . . .

a b s u r d i t y

Character Convolution (CharCNN)

$$\mathbf{h}[1] = \tanh(\mathbf{C}[*, 1:3] \otimes \mathbf{H} + b)$$

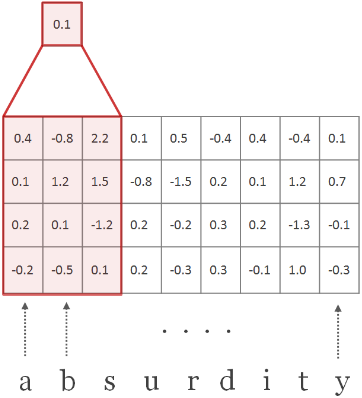
0.4	-0.8	2.2	0.1	0.5	-0.4	0.4	-0.4	0.1
0.1	1.2	1.5	-0.8	-1.5	0.2	0.1	1.2	0.7
0.2	0.1	-1.2	0.2	-0.2	0.3	0.2	-1.3	-0.1
-0.2	-0.5	0.1	0.2	-0.3	0.3	-0.1	1.0	-0.3

. . . .

a b s u r d i t y

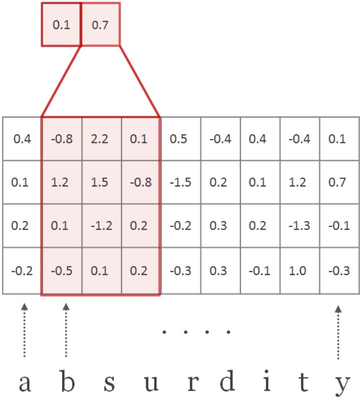
Character Convolution (CharCNN)

$$\mathbf{h}[1] = \tanh(\mathbf{C}[*, 1:3] \otimes \mathbf{H} + b)$$



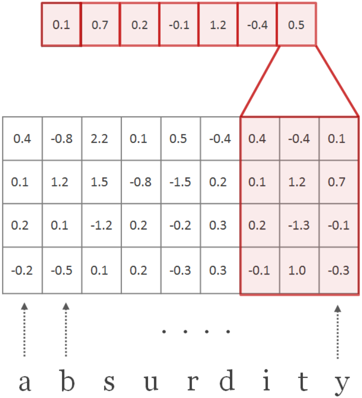
Character Convolution (CharCNN)

$$\mathbf{h}[2] = \tanh(\mathbf{C}[*, 2:4] \otimes \mathbf{H} + b)$$



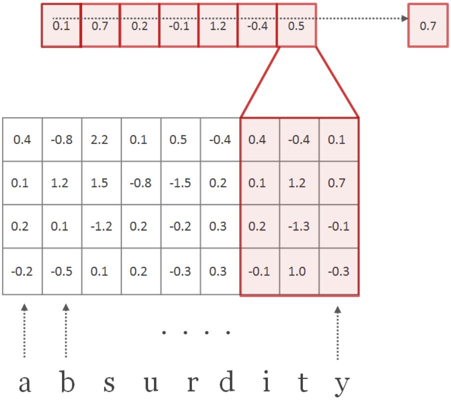
Character Convolution (CharCNN)

$$\mathbf{h}[T-2] = \mathrm{tanh}(\mathbf{C}[*, T-2:T] \otimes \mathbf{H} + b)$$



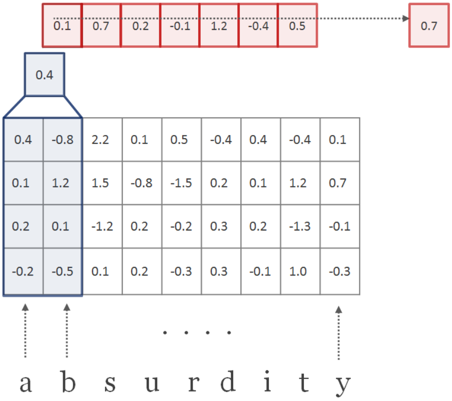
Character Convolution (CharCNN)

$$y[1] = \max_i \mathbf{h}[i]$$



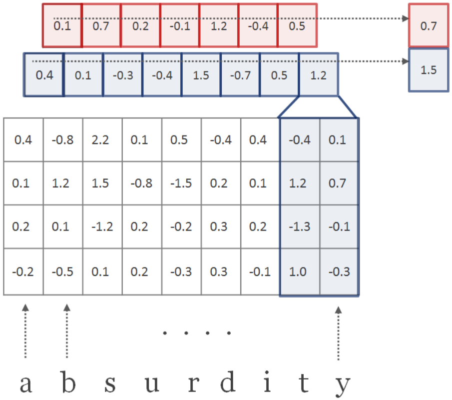
Character Convolution (CharCNN)

$$\mathbf{h}'[1] = \tanh(\mathbf{C}[*, 1:2] \otimes \mathbf{H}' + b')$$

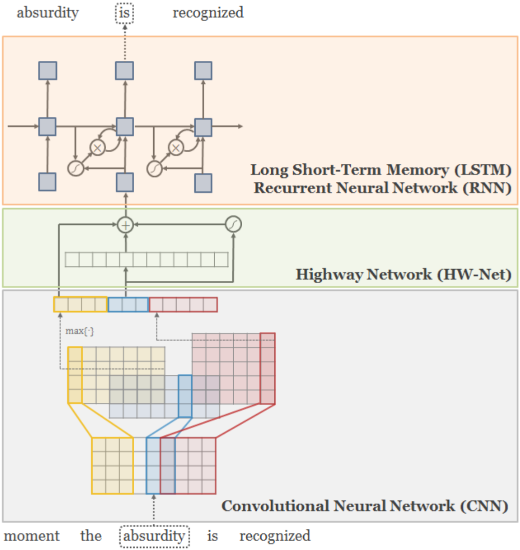


Character Convolution (CharCNN)

$$y[2] = \max_i \mathbf{h}'[i]$$



Convolution into Recurrent Model



Results: English PTB

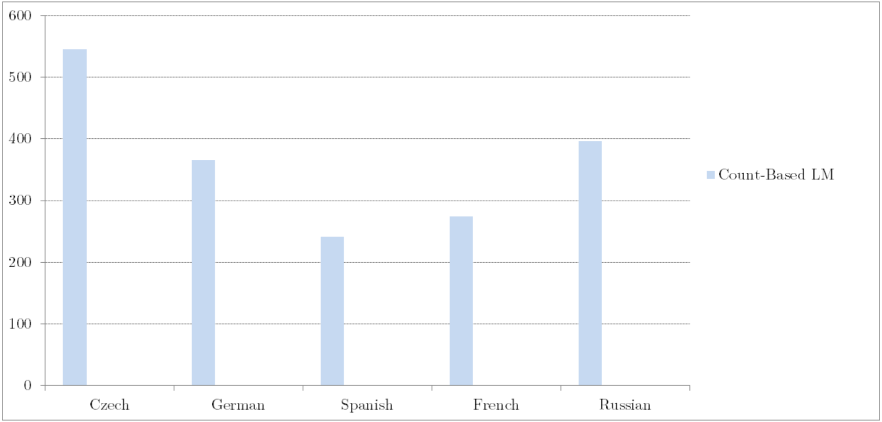
	Perplexity	Param Size
LSTM-Word-Small	97.6	5 M
LSTM-CharCNN-Small	92.3	5 M
LSTM-Word-Large	85.4	20 M
LSTM-CharCNN-Large	78.9	19 M
KN-5 (Mikolov et al. 2012)	141.2	2 M
RNN (Mikolov et al. 2012)	124.7	6 M
LSTM-Medium (Zaremba et al. 2014)	82.7	20 M
LSTM-Huge (Zaremba et al. 2014)	78.4	52 M

Data

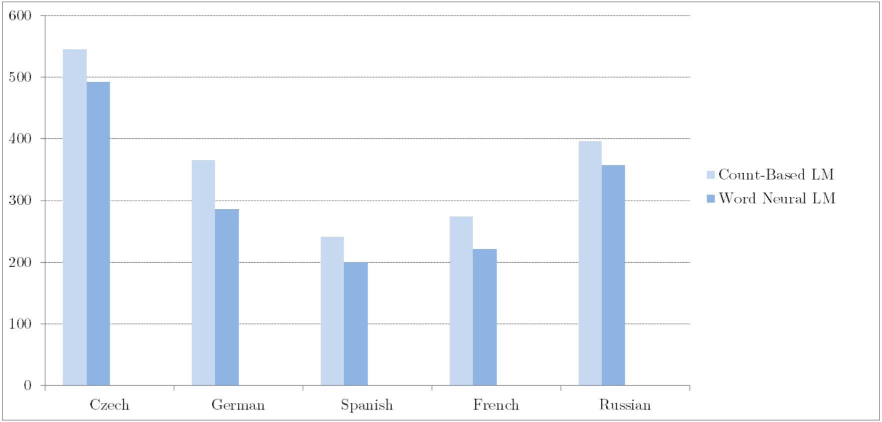
	Data-s			Data-l		
	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{C}\rvert$	T	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{C}\rvert$	T
English (En)	10 k	51	1 m	60 k	129	20 m
Czech (Cs)	46 k	93	1 m	206 k	127	17 m
German (De)	36 k	75	1 m	339 k	140	51 m
Spanish (Es)	27 k	72	1 m	152 k	130	56 m
French (Fr)	25 k	77	1 m	137 k	133	57 m
Russian (Ru)	62 k	64	1 m	497 k	114	25 m

Small English data is the English Penn Treebank (PTB). Rest comes from the 2013 ACL Workshop on Machine Translation.

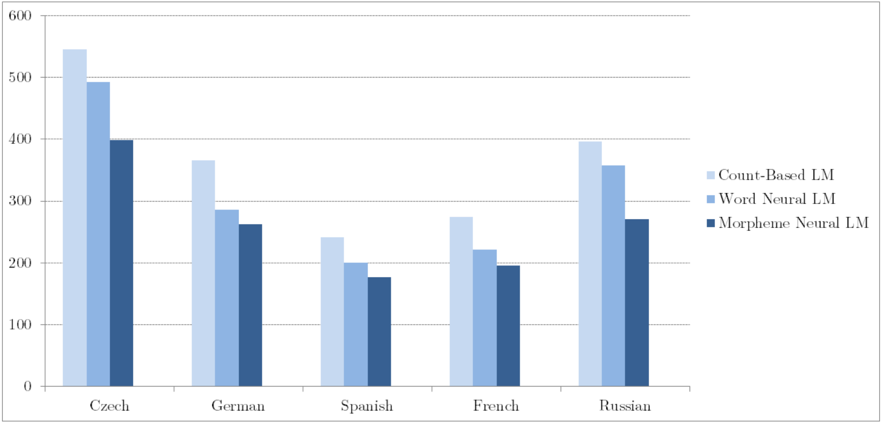
Results: Perplexity in Various Languages



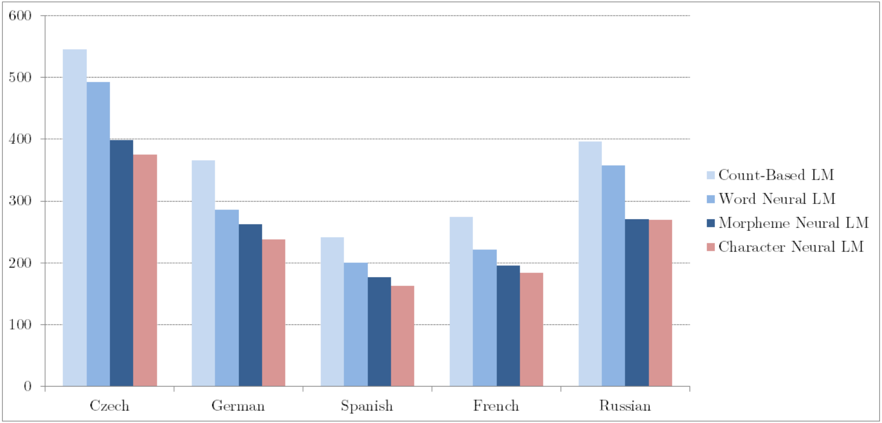
Results: Perplexity in Various Languages



Results: Perplexity in Various Languages



Results: Perplexity in Various Languages



Results: Large Datasets

		Cs	De	Es	Fr	Ru	En
B&B	KN-4	862	463	219	243	390	291
	MLBL	643	404	203	227	**300**	273
Small	Word	701	347	186	202	353	236
	Morph	615	331	189	209	331	233
	Char	**587**	**298**	**168**	**191**	313	**214**

Discussion: Learned Word Embeddings

	In Vocabulary				
	while	*his*	*you*	*richard*	*trading*
	although	your	conservatives	jonathan	advertised
LSTM	letting	her	we	robert	advertising
	though	my	guys	neil	turnover
	whole	this	your	gerard	training
LSTM-CNN	though	their	doug	edward	traded
	nevertheless	your	i	carl	traderg

Discussion: Learned Word Embeddings

	Out-of-Vocabulary		
	computer-aided	*misinformed*	*looooook*
LSTM-CharCNN	computer-guided	informed	look
	computer-driven	performed	looks
	computerized	outperformed	looked
	computer	transformed	looking

Abstractive Sentence Summarization (Rush et al., 2015)

Sentence Summarization

Source

Russian Defense Minister Ivanov called Sunday for the creation of a joint front for combating global terrorism.

Target

Russia calls for joint front against terrorism.

Summarization Phenomena:

- Generalization
- Deletion
- Paraphrase

Sentence Summarization

Source

Russian Defense Minister Ivanov *called Sunday for the creation of a joint front for combating global terrorism.*

Target

Russia *calls for joint front against terrorism.*

Summarization Phenomena:

- **Generalization**
- Deletion
- Paraphrase

Sentence Summarization

Source

Russian Defense Minister Ivanov called **Sunday** *for the creation of a joint front for combating global terrorism.*

Target

Russia calls for joint front against terrorism.

Summarization Phenomena:

- Generalization
- **Deletion**
- Paraphrase

Sentence Summarization

Source

Russian Defense Minister Ivanov called Sunday for the creation of a joint front **for combating** *global terrorism.*

Target

Russia calls for joint front **against** *terrorism.*

Summarization Phenomena:

- Generalization
- Deletion
- **Paraphrase**

Elements of Human Summary
Jing (2002)

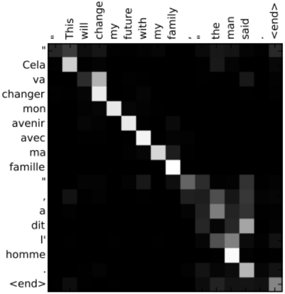
	Phenomenon	Abstract	Compress	Extract
(1)	Sentence Reduction	✓	✓	✓
(2)	Sentence Combination	✓	✓	✓
(3)	Syntactic Transformation	✓		✓
(4)	Lexical Paraphrasing	✓		
(5)	Generalization or Specification	✓		
(6)	Reordering	✓		✓

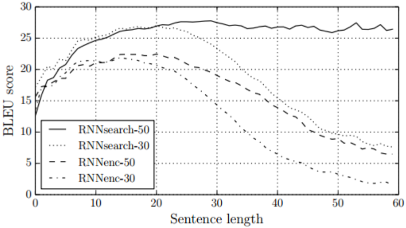
Related Work: Ext/Abs Sentence Summary

- **Syntax-Based** (Dorr et al., 2003; Cohn and Lapata, 2008; Woodsend et al., 2010)
- **Topic-Based** (Zajic et al., 2004)
- **Machine Translation-Based** (Banko et al., 2000)
- **Semantics-Based** (Liu et al., 2015)

Related Work: Attention-Based Neural MT
(Bahdanau et al., 2014)

- Use attention ("soft alignment") over source to determine next word.
- Robust to longer sentences versus encoder-decoder style models.
- No explicit alignment step, trained end-to-end.





Attention-Based Summarization (ABS)

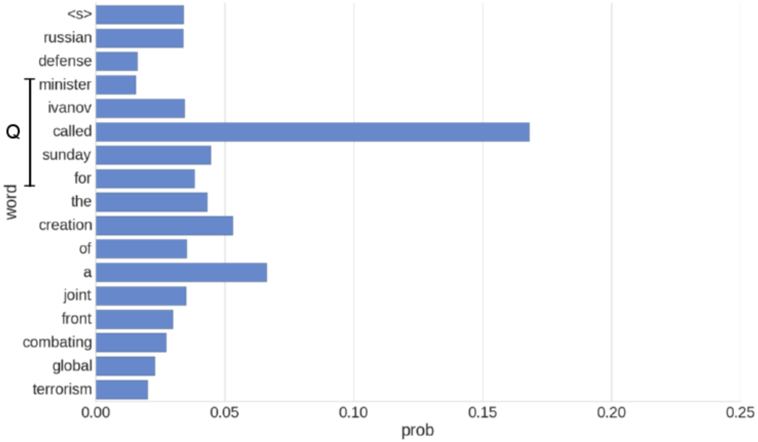
- $\mathbf{x}$; Source sentence of length M with $M >> N$
- $\mathbf{w}$; Summarized sentence of length N (we assume N is given)

$$
\begin{aligned}
\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\
\tilde{\mathbf{w}}'_{\mathrm{c}} &= [\mathbf{G}\mathbf{w}_{i-C+1}, \ldots, \mathbf{G}\mathbf{w}_i], \\
\mathbf{p} &\propto \exp(\tilde{\mathbf{x}}\mathbf{P}\tilde{\mathbf{w}}'_{\mathrm{c}}), \quad \textbf{[Attention Distribution]} \\
\forall i \quad \bar{\mathbf{x}}_i &= \sum_{q=i-(Q-1)/2}^{i+(Q-1)/2} \tilde{\mathbf{x}}_i / Q, \quad \textbf{[Local Smoothing]} \\
\mathrm{src}_3(\mathbf{x}, \mathbf{w}_{\mathrm{c}}) &= \mathbf{p}^{\top}\bar{\mathbf{x}}.
\end{aligned}
$$

ABS Example

$[\langle s \rangle$ Russia calls$]$	**for**
$\mathbf{w}_c$	$\mathbf{w}_{i+1}$

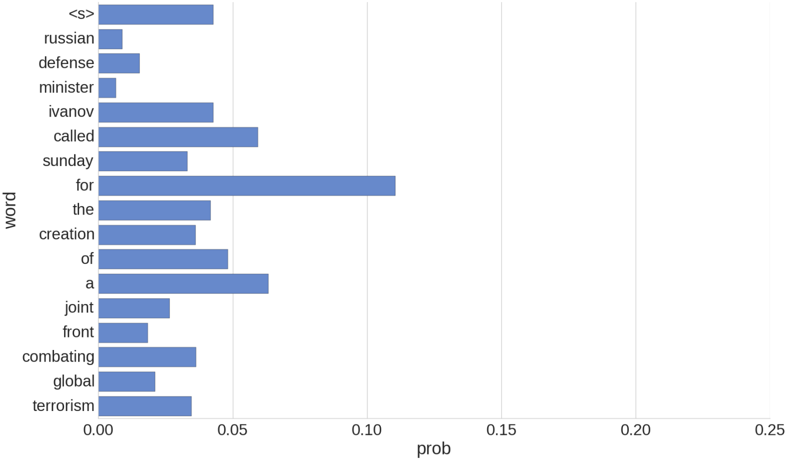
$\mathbf{x}$



ABS Example

$[\langle s \rangle$ Russia calls for$]$	**joint**
$\mathbf{w}_c$	$\mathbf{w}_{i+1}$

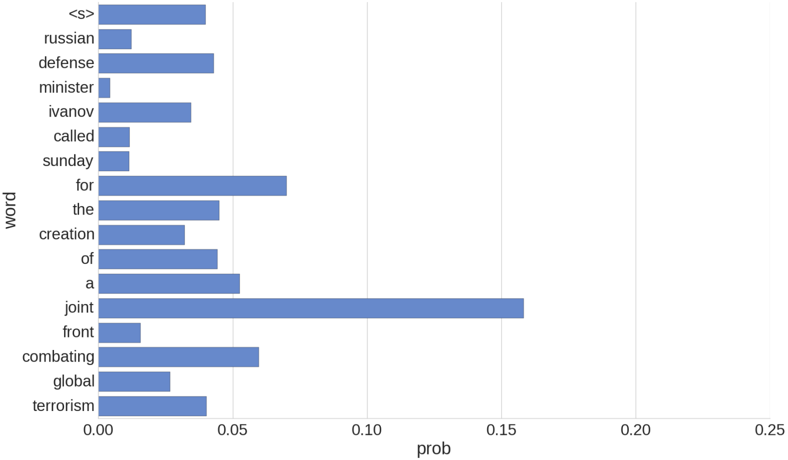
$\mathbf{x}$



ABS Example

[⟨s⟩ Russia calls for joint]	**front**
$\mathbf{w}_c$	$\mathbf{w}_{i+1}$

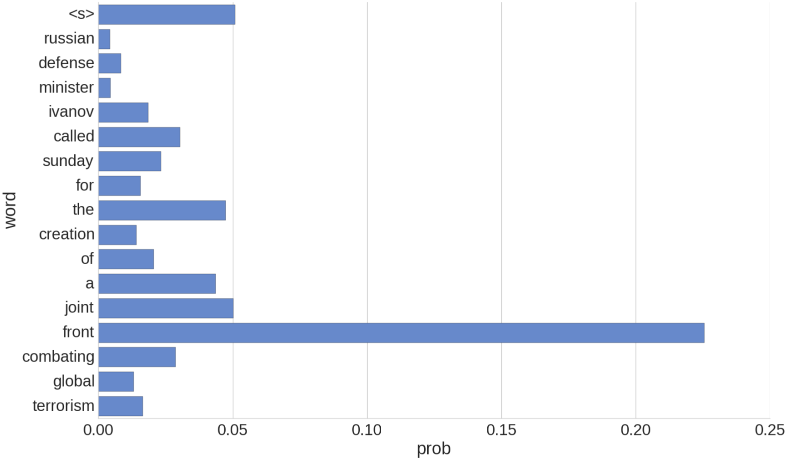
$\mathbf{x}$



ABS Example

⟨s⟩	[Russia calls for joint front]	**against**
	$\mathbf{w}_c$	$\mathbf{w}_{i+1}$

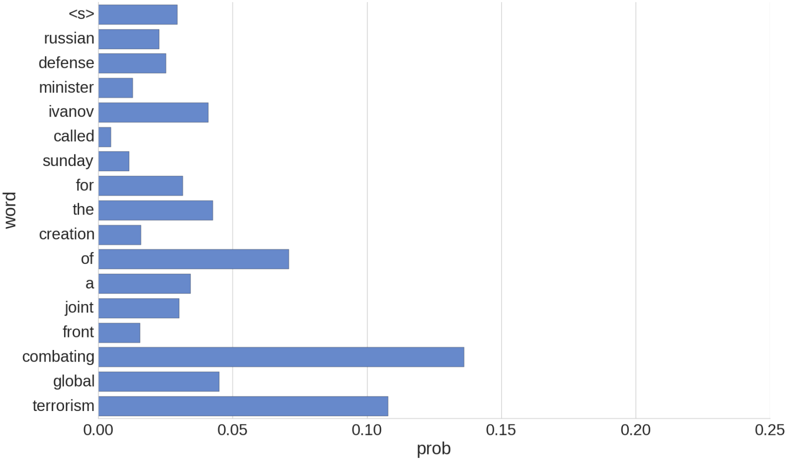
$\mathbf{x}$



ABS Example

⟨s⟩ Russia	[calls for joint front against]		**terrorism**
	$\mathbf{w}_c$		$\mathbf{w}_{i+1}$

$\mathbf{x}$

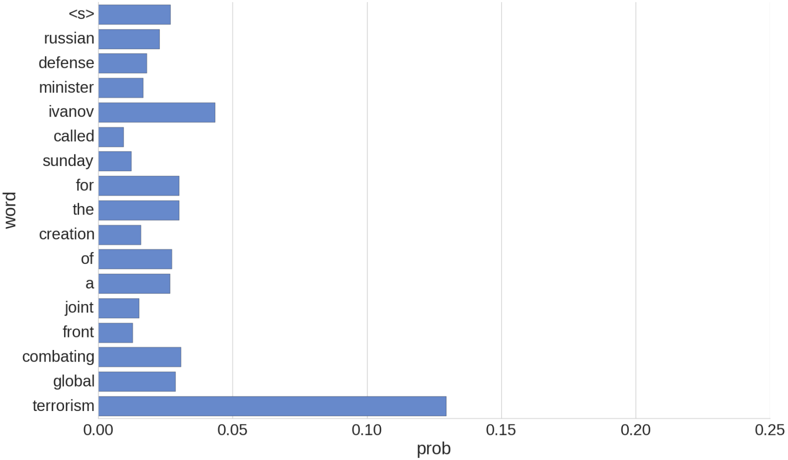


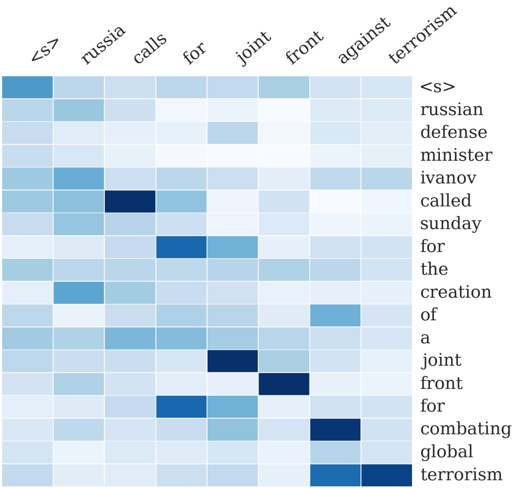
ABS Example

$\langle$s$\rangle$ Russia calls [for joint front against terrorism] .

$\mathbf{w}_c$ $\mathbf{w}_{i+1}$

$\mathbf{x}$





<s>
russia
calls
for
joint
front
against
terrorism
<s>
russian
defense
minister
ivanov
called
sunday
for
the
creation
of
a
joint
front
for
combating
global
terrorism

Sep 13, 3:17 PM EDT

GERMANY IMPLEMENTS TEMPORARY BORDER CHECKS TO LIMIT MIGRANTS

BY GEIR MOULSON AND SHAWN POGATCHNIK

ASSOCIATED PRESS

BERLIN (AP) -- Germany introduced temporary border controls Sunday to stem the tide of thousands of refugees streaming across its frontier, sending a clear message to its European partners that it needs more help with an influx that is straining its ability to cope.

AP Photo/Kay Nietfeld

Germany is a preferred destination for many people fleeing Syria's civil war and other troubled nations in the migration crisis that has bitterly divided Europe. They have braved dangerous sea crossings in flimsy

Headline Generation Training Set

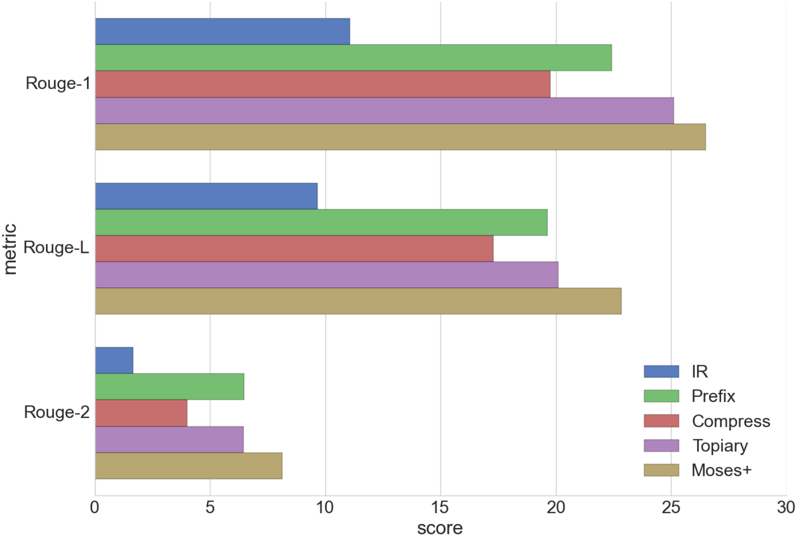
(Graff et al., 2003; Napoles et al., 2012)

- Use Gigaword dataset.

Total Sentences	3.8 M
Newswire Services	7
Source Word Tokens	119 M
Source Word Types	110 K
Average Source Length	31.3 tokens
Summary Word Tokens	31 M
Summary Word Types	69 K
Average Summary Length	8.3 tokens
Average Overlap	4.6 tokens
Average Overlap in first 75	2.6 tokens

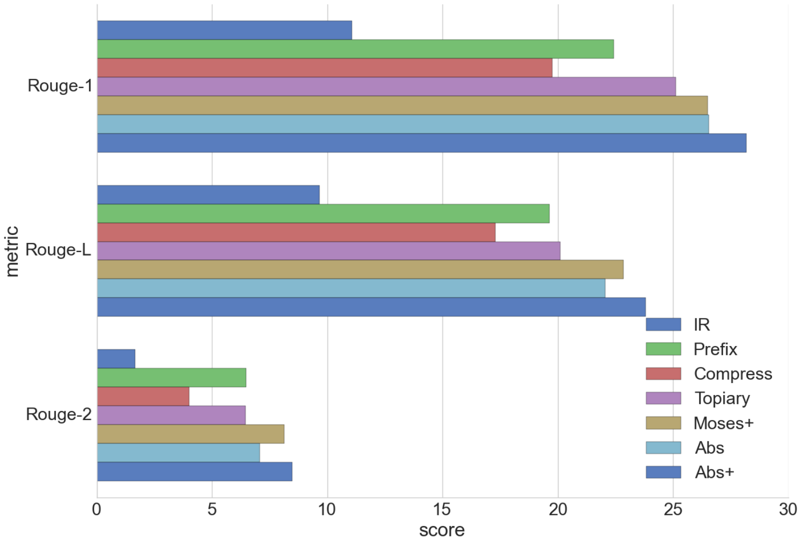
Summarization Results: DUC 2004

(500 pairs, 4 references, 75 characters)



Summarization Results: DUC 2004

(500 pairs, 4 references, 75 characters)



Generated Sentences on Gigaword I

Source:

a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted , a to p judiciary official said tuesday .

Ref: iranian-american academic held in tehran released on bail

Abs: detained iranian-american academic released from jail after posting bail

Generated Sentences on Gigaword II

Source:

ministers from the european union and its mediterranean neighbors gathered here under heavy security on monday for an unprecedented conference on economic and political cooperation .

Ref: european mediterranean ministers gather for landmark conference by julie bradford

Abs: mediterranean neighbors gather for unprecedented conference **on heavy security**

Generated Sentences on Gigaword III

Source:

the death toll from a school collapse in a haitian shanty-town rose to ## after rescue workers uncovered a classroom with ## dead students and their teacher , officials said saturday .

Ref: toll rises to ## in haiti school unk : official

Abs: death toll in haiti school **accident** rises to ##

Generated Sentences on Gigaword IV

Source:

australian foreign minister stephen smith sunday congratulated new zealand 's new prime minister-elect john key as he praised ousted leader helen clark as a " gutsy " and respected politician .

Ref: time caught up with nz 's gutsy clark says australian fm

Abs: australian foreign minister congratulates **new nz pm after election**

Generated Sentences on Gigaword V

Source:

two drunken south african fans hurled racist abuse at the country 's rugby sevens coach after the team were eliminated from the weekend 's hong kong tournament , reports said tuesday .

Ref: rugby union : racist taunts mar hong kong sevens : report

Abs: south african fans hurl racist **taunts** **at rugby sevens**

Generated Sentences on Gigaword VI

Source:

christian conservatives – kingmakers in the last two us presidential elections – may have less success in getting their pick elected in #### , political observers say .

Ref: christian conservatives power diminished ahead of #### vote

Abs: christian conservatives may have less success in #### election

Generated Sentences on Gigaword VII

Source:

the white house on thursday warned iran of possible new sanctions after the un nuclear watchdog reported that tehran had begun sensitive nuclear work at a key site in defiance of un resolutions .

Ref: us warns iran of step backward on nuclear issue

Abs: **iran** warns of possible new sanctions on nuclear work

Generated Sentences on Gigaword VIII

Source:

thousands of kashmiris chanting pro-pakistan slogans on sunday attended a rally to welcome back a hardline separatist leader who underwent cancer treatment in mumbai .

Ref: thousands attend rally for kashmir hardliner

Abs: thousands **rally in support** of hardline **kashmiri** separatist leader

Generated Sentences on Gigaword IX

Source:

an explosion in iraq 's restive northeastern province of diyala killed two us soldiers and wounded two more , the military reported monday .

Ref: two us soldiers killed in iraq blast december toll ###

Abs: # us two soldiers killed in restive northeast province

Generated Sentences on Gigaword X

Source:

russian world no. # nikolay davydenko became the fifth withdrawal through injury or illness at the sydney international wednesday , retiring from his second round match with a foot injury .

Ref: tennis : davydenko pulls out of sydney with injury

Abs: davydenko **pulls out** of sydney international with foot injury

Generated Sentences on Gigaword XI

Source:

russia 's gas and oil giant gazprom and us oil major chevron have set up a joint venture based in resource-rich northwestern siberia , the interfax news agency reported thursday quoting gazprom officials .

Ref: gazprom chevron set up joint venture

Abs: **russian oil giant chevron** set up **siberia joint venture**

Coreference Resolution (Wiseman et al., 2015)

Coreference Resolution

Cadillac posted a 3.2% increase despite new competition from Lexus, the fledgling luxury-car division of Toyota Motor Corp. Lexus sales weren't available; the cars are imported and Toyota reports their sales only at month-end.

Coreference Resolution

[Cadillac] posted a [3.2% increase] despite [new competition from [Lexus, the fledgling luxury-car division of [Toyota Motor Corp]]]. [[Lexus] sales] weren't available; [the cars] are imported and [Toyota] reports [[their] sales] only at [month-end].

Mention Ranking

(Denis and Baldridge, 2008; Bengtson and Roth, 2008)

- Model each mention x as having a single "true" antecedent
- Score potential antecedents y of each mention x with a scoring function $s(x, y)$
- $\mathcal{Y}(x) = \{\text{mentions before } x\} \cup \{\epsilon\}$
- Predict $y^* = \arg\max_{y \in \mathcal{Y}(x)} s(x, y)$

$s(x, y_1) = 1.2$ $s(x, y_2) = 0.9$ $s(x, \epsilon) =$ -1.8

. . . [the cars] are imported and [Toyota] reports [their]

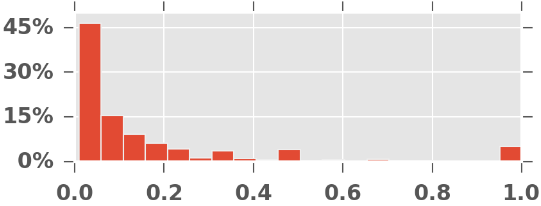
$\boxed{y_1}$ $\boxed{y_2}$ $\boxed{x}$

Simple Features Not Discriminative

E.g., is [Lexus sales] the antecedent of [their sales]?

- Common antecedent features: String/Head Match, Sentences Between, Mention-Antecedent Numbers/Heads/Genders, etc.

$$\phi_{\mathrm{p}}(\text{[their sales]},\text{[Lexus sales]}) = \left\{ \begin{array}{c} \text{string-match=false} \\ \text{head-match=true} \\ \text{sentences-between=0} \\ \text{ment-ant-numbers=plur.,plur.} \\ \vdots \end{array} \right\}$$



Dealing with the Feature Problem

Finding discriminative features a major challenge for coreference systems (Fernandes et al., 2012; Durrett and Klein, 2013)

- Typical to define (or search for) feature conjunction-schemes to improve predictive performance (Fernandes et al., 2012; Durrett and Klein, 2013; Björkelund and Kuhn, 2014).
- Not just a problem for Mention Ranking systems.

Extending the Piecewise Model I

Goal: learn higher order feature representations

We first define the following feature representations:

$$\boldsymbol{h}_{\mathrm{a}}(x) \triangleq \tanh(\mathbf{U}_{\mathrm{a}}^{\top}\, \boldsymbol{\phi}_{\mathrm{a}}(x) + \boldsymbol{b}_{\mathrm{a}})$$
$$\boldsymbol{h}_{\mathrm{p}}(x, y) \triangleq \tanh(\mathbf{U}_{\mathrm{p}}^{\top}\, \boldsymbol{\phi}_{\mathrm{p}}(x, y) + \boldsymbol{b}_{\mathrm{p}})$$

- **Here, $\phi_{\mathrm{a}}, \phi_{\mathrm{p}}$ are raw features.**
-

$$\mathbf{U}_{\textsf{ment-ant-numbers=plur.,plur.}}$$

$$\mathbf{U}_{\textsf{head-match=true}}$$

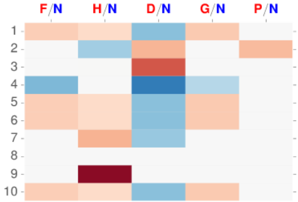
, etc.

Extending the Piecewise Model II

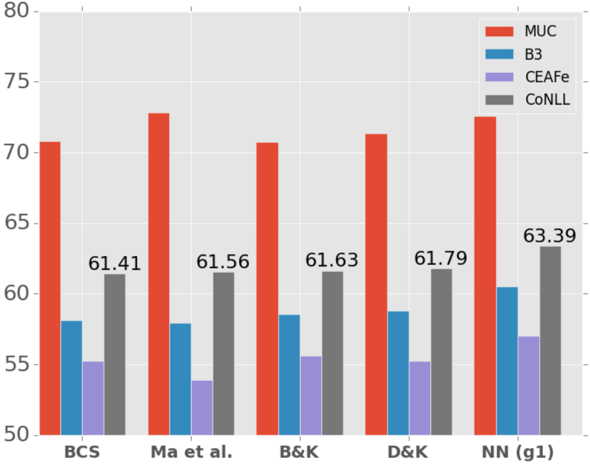
Use the scoring function

$$s(x,y) \triangleq \begin{cases} \boldsymbol{u}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{h}_{\mathrm{a}}(x) \\ \boldsymbol{h}_{\mathrm{p}}(x,y) \end{bmatrix} + u_0 & \text{if } y \neq \epsilon \\ \boldsymbol{v}^{\mathsf{T}} \boldsymbol{h}_{\mathrm{a}}(x) + v_0 & \text{if } y = \epsilon \end{cases}$$

Scoring function uses learned representations, for instance $\boldsymbol{h}_{\mathrm{p}}$:



Main Results



Results on CoNLL 2012 English test set. We compare with (in order) Durrett and Klein (2013), Ma et al. (2014), Björkelund and Kuhn (2014), and Durrett and Klein (2014). F_1 gains are significant ($p < 0.05$) compared with both B&K and D&K for all metrics.

Discussion: What are we getting wrong?

	Singleton		1^{st} in clust.		Anaphoric	
	FL	#	FL	#	FNWL	#
Ment. w/ prev. head match	817	8.2K	147	0.8K	700318	4.7K
Ment. w/o prev. head match	86	19.8K	41	2.4K	67759	1.0K
Pronominal mentions	948	2.6K	257	0.5K	434875	7.3K

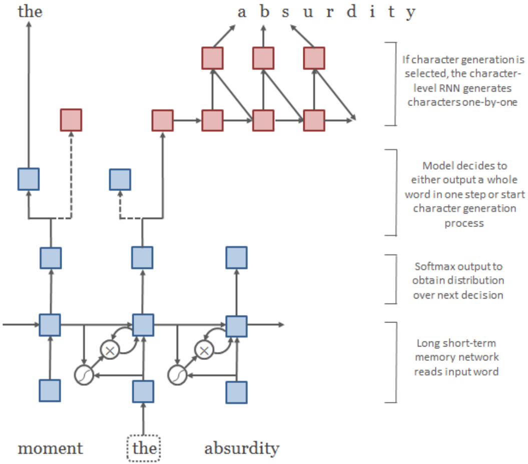
Largest % error on anaphoric mentions with no previous head match

- The classic "hard" coreference case, presumably requiring knowledge, understanding

But make most errors (by far) on pronouns!

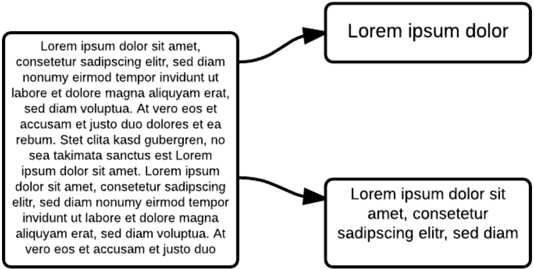
Future Directions: Character-Aware

Generating characters and machine translation.



Future Directions: Summarization

Complete Document Summarization



Future Directions: Coreference Resolution

Incorporating Document Context

$$
\begin{tikzcd}
 & & y_1 & & & & y_2 \\
h_1 \arrow[r] & h_2 \arrow[r] & h_3 \arrow[r] \arrow[u] & h_4 \arrow[r] & h_5 \arrow[r] & h_6 \arrow[r] & h_7 \arrow[u] \\
\text{forgotten} \arrow[u] & \text{[her} \arrow[u] & \text{badge]} \arrow[u] & \text{.} \arrow[u] & \text{Fortunately} \arrow[u] & \text{[Mr.} \arrow[u] & \text{Gordon]} \arrow[u]
\end{tikzcd}
$$

References I

Alexandrescu, A. and Kirchhoff, K. (2006). Factored neural language models. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pages 1–4. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

Banko, M., Mittal, V. O., and Witbrock, M. J. (2000). Headline generation based on statistical translation. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pages 318–325. Association for Computational Linguistics.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. The Journal of Machine Learning Research, 3:1137–1155.

References II

Bengtson, E. and Roth, D. (2008). Understanding the Value of Features for Coreference Resolution. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 294–303. Association for Computational Linguistics.

Bilmes, J. A. and Kirchhoff, K. (2003). Factored language models and generalized parallel backoff. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2, pages 4–6. Association for Computational Linguistics.

Björkelund, A. and Kuhn, J. (2014). Learning structured perceptrons for coreference Resolution with Latent Antecedents and Non-local Features. ACL, Baltimore, MD, USA, June.

References III

Botha, J. and Blunsom, P. (2014). Compositional Morphology for Word Representations and Language Modelling. In Proceedings of ICML.

Chatterjee, P., Mukherjee, S., Chaudhuri, S., and Seetharaman, G. (2009). Application of papoulis–gerchberg method in image super-resolution and inpainting. The computer journal, 52(1):80–89.

Cohn, T. and Lapata, M. (2008). Sentence compression beyond word deletion. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pages 137–144. Association for Computational Linguistics.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. Journal of Machine Learning Research, 12:2493–2537.

References IV

Denis, P. and Baldridge, J. (2008). Specialized Models and Ranking for Coreference Resolution. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 660–669. Association for Computational Linguistics.

Dorr, B., Zajic, D., and Schwartz, R. (2003). Hedge trimmer: A parse-and-trim approach to headline generation. In Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5, pages 1–8. Association for Computational Linguistics.

Durrett, G. and Klein, D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1971–1982.

Durrett, G. and Klein, D. (2014). A Joint Model for Entity Analysis: Coreference, Typing, and Linking. Transactions of the Association for Computational Linguistics, 2:477–490.

References V

Fernandes, E. R., Dos Santos, C. N., and Milidiú, R. L. (2012). Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In Joint Conference on EMNLP and CoNLL-Shared Task, pages 41–48. Association for Computational Linguistics.

Graff, D., Kong, J., Chen, K., and Maeda, K. (2003). English gigaword. Linguistic Data Consortium, Philadelphia.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9:1735–1780.

Jing, H. (2002). Using hidden markov modeling to decompose human-written summaries. Computational linguistics, 28(4):527–543.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. In Proceedings of ACL 2014.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In Proceedings of EMNLP 2014.

References VI

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2015). Character-Aware Neural Language Models. ArXiv e-prints.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Handwritten Digit Recognition with a Backpropagation Network. In Proceedings of NIPS.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., and Smith, N. A. (2015). Toward abstractive summarization using semantic representations.

Luong, M.-T., Socher, R., and Manning, C. (2013). Better Word Representations with Recursive Neural Networks for Morphology. In Proceedings of CoNLL.

Ma, C., Doppa, J. R., Orr, J. W., Mannem, P., Fern, X., Dietterich, T., and Tadepalli, P. (2014). Prune-and-score: Learning for Greedy Coreference Resolution. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.

References VII

Marton, Y., Habash, N., and Rambow, O. (2010). Improving arabic dependency parsing with lexical and inflectional morphological features. In Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 13–21. Association for Computational Linguistics.

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., and Khudanpur, S. (2010). Recurrent Neural Network Based Language Model. In Proceedings of INTERSPEECH.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated gigaword. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pages 95–100. Association for Computational Linguistics.

References VIII

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In To appear in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, 18-21 September 2015, Lisbon, Portugal.

Wiseman, S., Rush, A. M., Shieber, S. M., and Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pages 1416–1426.

References IX

Woodsend, K., Feng, Y., and Lapata, M. (2010). Generation with quasi-synchronous grammar. In Proceedings of the 2010 conference on empirical methods in natural language processing, pages 513–523. Association for Computational Linguistics.

Zajic, D., Dorr, B., and Schwartz, R. (2004). Bbn/umd at duc-2004: Topiary. In Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston, pages 112–119.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent Neural Network Regularization. arXiv:1409.2329.