Machine Translation: Going Deep

Philipp Koehn

4 June 2015





• More data

• Better linguistically motivated models

• Better machine learning



• More data

• Better linguistically motivated models

• Better machine learning



what problems do we need to solve?

Word Translation Problems



• Words are ambiguous

He deposited money in a bank account with a high interest rate.

Sitting on the bank of the Mississippi, a passing ship piqued his interest.

- How do we find the right meaning, and thus translation?
- Context should be helpful

Phrase Translation Problems



• Idiomatic phrases are not compositional

It's raining cats and dogs.

Es schüttet aus Eimern.

(it pours from buckets.)

• How can we translate such larger units?

Syntactic Translation Problems



• Languages have different sentence structure

das	behaupten	sie	wenigstens	
this	claim	they	at least	
the		she		

- Convert from object-verb-subject (OVS) to subject-verb-object (SVO)
- Ambiguities can be resolved through syntactic analysis
 - the meaning the of das not possible (not a noun phrase)
 - the meaning she of sie not possible (subject-verb agreement)

Semantic Translation Problems



• Pronominal anaphora

I saw the movie and it is good.

- How to translate it into German (or French)?
 - it refers to movie
 - movie translates to Film
 - Film has masculine gender
 - ergo: it must be translated into masculine pronoun er
- We are not handling this very well [Le Nagard and Koehn, 2010]

Semantic Translation Problems



• Coreference

Whenever I visit my uncle and his daughters, I can't decide who is my favorite cousin.

- How to translate cousin into German? Male or female?
- Complex inference required

Discourse Translation Problems



• Discourse

Since you brought it up, I do not agree with you.

Since you brought it up, we have been working on it.

- How to translated since? Temporal or conditional?
- Analysis of discourse structure a hard problem

Mismatch in Information Structure



- Morphology allows adding subtle or redundant meaning
 - verb tenses: time action is occurring, if still ongoing, etc.
 - count (singular, plural): how many instances of an object are involved
 - definiteness (the cat vs. a cat): relation to previously mentioned objects
 - grammatical gender: helps with co-reference and other disambiguation
- Some languages allow repeated information across sentences to be dropped
 - 1. Yesterday Jane bought an apple in the store.
 - 2. Ate.



linguistically motivated models

Synchronous Grammar Rules



• Nonterminal rules

 $\mathsf{NP} \to \mathsf{DET}_1 \: \mathsf{NN}_2 \: \mathsf{JJ}_3 \mid \mathsf{DET}_1 \: \mathsf{JJ}_3 \: \mathsf{NN}_2$

• Terminal rules

 $N \rightarrow maison \mid house$ $NP \rightarrow la maison bleue \mid the blue house$

• Mixed rules

 $NP \rightarrow la \text{ maison } JJ_1 \mid \text{ the } JJ_1 \text{ house}$

Learning Rules [GHKM]





Extracted rule: $VP \rightarrow X_1 X_2$ aushändigen | passing on $PP_1 NP_2$

Syntactic Decoding



Inspired by monolingual syntactic chart parsing:

During decoding of the source sentence, a chart with translations for the ${\cal O}(n^2)$ spans has to be filled







German input sentence with tree





Purely lexical rule: filling a span with a translation (a constituent in the chart)





Purely lexical rule: filling a span with a translation (a constituent in the chart)





Purely lexical rule: filling a span with a translation (a constituent in the chart)





Complex rule: matching underlying constituent spans, and covering words





Complex rule with reordering







- Chart consists of cells that cover contiguous spans over the input sentence
- Each cell contains a set of hypotheses
- Hypotheses are constructed bottom-up
- Various ways to binarize rules we use CKY+

Feature Structures



- Various forms of long distance agreement
 - subject-verb in count (president agrees vs. presidents agree)
 - subject-verb in person (he says vs. I say)
 - verb subcategorization
 - noun phrases in gender, case, count (a big house vs. big houses)
- Represent syntactic constituents with feature structures

CAT	np -
HEAD	house
CASE	subject
COUNT	plural
PERSON	3rd

Constraints



• Grammar rules may be associated with constraints

$S \to NP \; VP$

S[head] = VP[head] NP[count] = VP[count] NP[person] = VP[person] NP[case] = subject

- Simpler: for each type of non-terminal (NP, VP, S) to be generated \rightarrow set of checks
- Used for
 - case agreement in noun phrases [Williams and Koehn, 2011]
 - consistent verb complex [Williams and Koehn, 2014]

State of the Art



• Good results for German–English [WMT 2014]

language pair	syntax preferred
German–English	57%
English–German	55%

• Mixed for other language pairs

language pair	syntax preferred
Czech–English	44%
Russian–English	44%
Hindi–English	54%

• Also very successful for Chinese–English

Results in 2015



• German–English

	2013	2014	2015
UEDIN phrase-based	26.8	28.0	29.3
UEDIN syntax	26.6	28.2	28.7
Δ	-0.2	+0.2	-0.6
Human preference	52%	57%	?

• English-German

	2013	2014	2015
UEDIN phrase-based	l 20.1	20.1	22.8
UEDIN syntax	19.4	20.1	24.0
Δ	-0.7	+0.0	+1.2
Human preference	55%	55%	?

Perspective



- Syntax-based models superior for German \leftrightarrow English
 - also previously shown for Chinese–English (ISI)
 - some evidence for low resource languages (Hindi)
- Next steps
 - Enforcing correct subcategorization frames
 - Features over syntactic dependents
 - Condition on source side syntax (soft features, rules, etc.)
- Decoding still a challenge
- Extend to AMRs?



a disruption: deep learning

Linear Models



• We used before weighted linear combination of feature values h_j and weights λ_j

$$\operatorname{score}(\lambda, \mathbf{d}_i) = \sum_j \lambda_j \ h_j(\mathbf{d}_i)$$

• Such models can be illustrated as a "network"



Limits of Linearity



- We can give each feature a weight
- But not more complex value relationships, e.g,
 - any value in the range [0;5] is equally good
 - values over 8 are bad
 - higher than 10 is not worse



• Linear models cannot model XOR



Multiple Layers



• Add an intermediate ("hidden") layer of processing (each arrow is a weight)



• Have we gained anything so far?

Non-Linearity



• Instead of computing a linear combination

$$\operatorname{score}(\lambda, \mathbf{d}_i) = \sum_j \lambda_j h_j(\mathbf{d}_i)$$

• Add a non-linear function

$$\operatorname{score}(\lambda, \mathbf{d}_i) = f\left(\sum_j \lambda_j \ h_j(\mathbf{d}_i)\right)$$

• Popular choices



(sigmoid is also called the "logistic function")

Deep Learning



• More layers = deep learning



I Told You So!



- My first publications
 - Combining Genetic Algorithms and Neural Networks Philipp Koehn, MSc thesis 1994
 - Genetic Encoding Strategies for Neural Networks Philipp Köhn, IPMU 1996
 - Combining Multiclass Maximum Entropy Text Classifiers with Neural Network Voting Philipp Koehn, PorTAL 2002
- Real credit goes to Holger Schwenk (continuous space language models for statistical machine translation in 2006)



- Words represented by 1-hot vector
- Map each word first into a lower-dimensional real-valued space
- One hidden layer
- Predict next word in output (1-hot vector)

Word Embeddings





- By-product: embedding of word into continuous space
- Similar contexts \rightarrow similar embedding
- Recall: distributional semantics

Word Embeddings



surrounding opposite oùtside		ry limi key	educed ted	total		bg ţtom		past NUMBE	mile BS decade yearth minute	rter half - round	head wi face an side dg hand c
acro≨fobøgether	forward offf dpwn straj	Similar close related	1		particular	standary	L	chaniations.	day (perio era	season spo d stage	t box screen
behind	away Joing apart back		0	open			electric			drama theater	press val d
	Ĩ	eft					mobile	e internet media	bass guita solo pi concerta	orchestra r opera ano style	band color
	, baving		arowina -lea	d		đaž	fm nLine { Ty entert	adio ElevisigNa1de607 AdVASUTng	si nic demosty	Musicart nging ′musicalaudien	sound nee voice image character
sent	have		developing					news d u ld talk	guest : new spa g	studiootilehow 9 MMAIne	series scene
	speaking st	sbareding co	supporting using mtaining prod	26]]j na a		_	live	r open	e Mhlish ing ing	hibut backgrout i	pl stouthok eltext song
li acting	ving	featured	cre pathowing performing 1,	Ating giving ting eaving	scon 105img ^{ing} greaching	riwg Dayring	bost		writing reading	speech fea fashion	episode ^{wa} ture nama title
med educated medained	mixed shared	a	₫ġįt¥i ng	foldin pass ruum:	ng ing ing		aired broadcast	, _]			reference
ed applie	charyed d	equivalent	urru up t exen ow	moving ing ng return	ing '	ru setti fiğt ending	hit	in the second	' display building meeting	release launch Flanding	tx: chave
based develop	com	tached Nected closed					ano Cas t			Liying -	cover turn
sefemded		coules apped	sta	mding	¢8€j¥# nin	g kill	ing	st.	isipt	tour	end start

Word Embeddings









- Morphosyntactic regularities (Mikolov et al., 2013)
 - adjectives base form vs. comparative, e.g., good, better
 - nouns singular vs. plural, e.g., year, years
 - verbs present tense vs. past tense, e.g., see, saw
- Semantic regularities
 - clothing is to shirt as dish is to bowl
 - evaluated on human judgment data of semantic similarities



machine translation with neural networks

Feed Forward Neural Network





Recurrent Neural Network



43

Encoder–Decoder Model



- Word embeddings seen as "semantic representations"
- Recurrent Neural Network
 → semantic representation of whole sentence
- Idea
 - encode semantics of the source sentence with recurrent neural network
 - decode semantics into target sentence from recurrent neural network
- Model $(w_1, ..., w_{l_f+l_e}) = (f_1, ..., f_{l_f}, e_1, ..., e_{l_e})$ $\prod_k p(w_1, ..., w_{l_f+l_e}) = \prod p(w_k | w_1, ..., w_{k-1})$
- But: bias towards end of sentence



LSTM and Reversed Order (Sutskever et al., 2014)

- Long short term memory for better retention of long distance memory
- Reverse production of target sentence

 $(f_1, ..., f_{l_f}, e_{l_e}, ..., e_1)$

- Some tricks (ensemble learning)
- Claims that it works as stand-alone model but better in reranking



Convolutional Neural Networks (Kalchbrenner and Blunsom, 2013)



- Build sentence representation bottom-up
 - merge any *n* neighboring nodes
 - *n* may be 2, 3, ...
- Generate target sentence by inverting the process
- Used successfully in re-ranking (Cho et al., 2014)



Adding an Alignment Model (Bahdanau, Cho and Bengio, 2015)

- Recurrent neural networks to create context representations for each input word
- Alignment model: conditioned on previous state and source side context



• Comment: this feels a bit like the HMM variant of the IBM Models

Does Any of This Work?



• Papers claim gains (sometimes only in reranking)

• Montreal (Bahdanau, Cho and Bengio, 2015) submission to WMT 2015

	de-en	en-de	cs-en	en-cs	fi-en
Best SMT	29.3	24.0	26.2	18.2	19.7
Montreal	27.9	22.4	23.8	18.4	13.6
Montreal emsemble		24.9			

(Scores from matrix.statmt.org)

Reflections



- Traditional statistical models have real short-comings
 - how to back-off to less context?
 - how to cluster information among words?
- Neural networks offer a more flexible way to condition on context
- Two strategies
 - Incremental strategy: replace statistical components with neural components
 - Leap forward strategy: start from scratch: neural machine translation



syntax-based machine translation with neural networks

Dependency Structure







- Predict from left children (up to 2)
- Example: $p(coffee|cup, drink, a, \epsilon)$

Statistical Model



• Probability distribution

p(word|parent, grand-parent, left-most-sibling, 2nd-left-most-sibling) for instance

 $p(\text{coffee}|\text{cup},\text{drink},\text{a},\epsilon)$

- Difficult to model
 - very sparse
 - no sharing of information between

 $p(\mathsf{coffee}|\mathsf{cup},\mathsf{drink},\mathsf{a},\epsilon)$

and

 $p(\mathsf{tea}|\mathsf{cup},\mathsf{drink},\mathsf{a},\epsilon)$

Neural Network Model



• Probability distribution

p(word|parent, grand-parent, left-most-sibling, 2nd-left-most-sibling) can be converted straightforward into a feed-forward neural network

- Words encoded with embeddings
- Empty slots modeled by average embedding over all words

Results



• Sennrich (2015)

System	Newstest 2013	Newstest 2014
Baseline	20.0	20.5
+NNLM	20.6	21.1
+neural dependency	20.9	21.6
+NNLM+neural dependency	21.0	21.8

• Official submissions to WMT 2015

System	BLEU
UEDIN syntax	22.6
UEDIN syntax with neural models	24.0

Caution: there were also other differences



what needs to be done?

Error Analysis: The Rules



- Given:
 - bilingual speaker
 - source sentence
 - machine translation output
 - (possibly reference translation)
- Step 1: Make minimal correction to create acceptable translation (fluent target language, correct meaning, may not be stylistically perfect)
- Step 2: Identify errors
- Error categories
 - qualitative
 - 1 error may cover mulitiple words
- A subjective process

Example: Simple Errors



- **SRC:** Es geht also um viel mehr als um Partikularinteressen des Herren Medau", so Pötzl.
- **REF:** It's therefore about a lot more than the individual interests of the Medau gentleman," he said.
- **TGT:** It is so much more than vested interests of Mr Medau," said Pötzl.

Corrected Target: It is about so much more than the vested interests of Mr Medau," Pötzl said.

Errors:	$\epsilon ightarrow about$	 missing preposition
	$\epsilon ightarrow the$	 missing determiner
	said	 reordering error: verb

Example: Muddle



- **SRC:** Die Polizei von Karratha beschuldigt einen 20-jhrigen Mann der Nichtbeachtung eines Haltesignals sowie rcksichtslosen Fahrens.
- **REF:** Karratha Police have charged a 20-year-old man with failing to stop and reckless driving.
- **TGT:** The police believe the failure of a 20-year-old man accused of Karratha signal and reckless driving.
- **Corrected Target:** The police of Karratha charged a 20-year-old man with failure to obey a signal and reckless driving.

This is a muddle, there is just too much wrong to categorize individual errors.



- German–English
- Syntax-based system (UEDIN)
- WMT 2015 test set
- Examined 100 sentences
- One judge: me
- Note: small scale just for this talk

Results



- 2.85 errors per sentence on average
- Distribution

Sentences with	Count
0 errors	16 sentences
1 error	18 sentences
2 errors	17 sentences
3 errors	17 sentences
more than 3 errors	32 sentences

Results



- Longest sentence with no error
 - Source: Der Oppositionspolitiker Imran Khan wirft Premier Sharif vor, bei der Parlamentswahl im Mai vergangenen Jahres betrogen zu haben.
 - Target: The opposition politician Imran Khan accuses Premier Sharif of having cheated in the parliamentary election in May of last year.
 - Has a complex subclause construction: accuses ... of having cheated

Major Error Categories



Count	Category	Count	Category
29	Wrong content word - noun	6	Wrong content word - phrasal verb
25	Wrong content word - verb	6	Added function word - determiner
22	Wrong function word - preposition	5	Unknown word - noun
21	Inflection - verb	5	Missing content word - adverb
14	Reordering: verb	5	Missing content word - noun
13	Reordering: adjunct	5	Inflection - noun
12	Missing function word - preposition	4	Reordering: NP
10	Missing content word - verb	3	Missing content word - adjective
9	Wrong function word - other	3	Inflection - wrong POS
9	Wrong content word - wrong POS	3	Casing
9	Added punctuation	2	Unknown word - verb
8	Muddle	2	Reordering: punctuation
8	Missing function word - connective	2	Reordering: noun
8	Added function word - preposition	2	Reordering: adverb
7	Missing punctuation	2	Missing function word - determiner
7	Wrong content word - adverb	2	Inflection - adverb



• Word sentence disambiguation

Count	Category
29	Wrong content word - noun
25	Wrong content word - verb
9	Wrong content word - wrong POS
7	Wrong content word - adverb
6	Wrong content word - phrasal verb

• Prepositions

Count	Category
22	Wrong function word - preposition
12	Missing function word - preposition
8	Added function word - preposition

• Reordering

Count	Category
14	Reordering: verb
13	Reordering: adjunct
4	Reordering: NP
2	Reordering: noun
2	Reordering: adverb

Note: much less of a problem than with phrase models

• Other issues with verbs

Count	Category
21	Inflection - verb
10	Missing content word - verb



Thank You



questions?