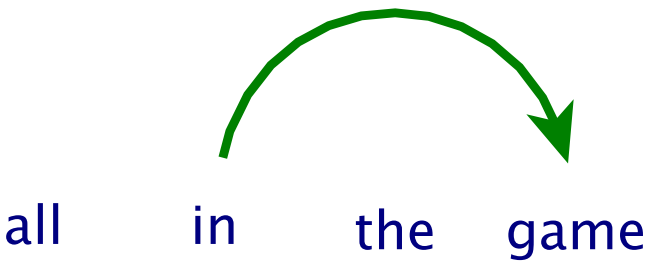
Natural Language Understanding in a Continuous Space

Karl-Moritz Hermann, Nal Kalchbrenner, Edward Grefenstette, and **Phil Blunsom**

phil.blunsom@cs.ox.ac.uk

Features and NLP



$$p(\text{game}|\text{in}) \propto \exp(\mathbf{w}^T \Phi(\text{game},\text{in}))$$

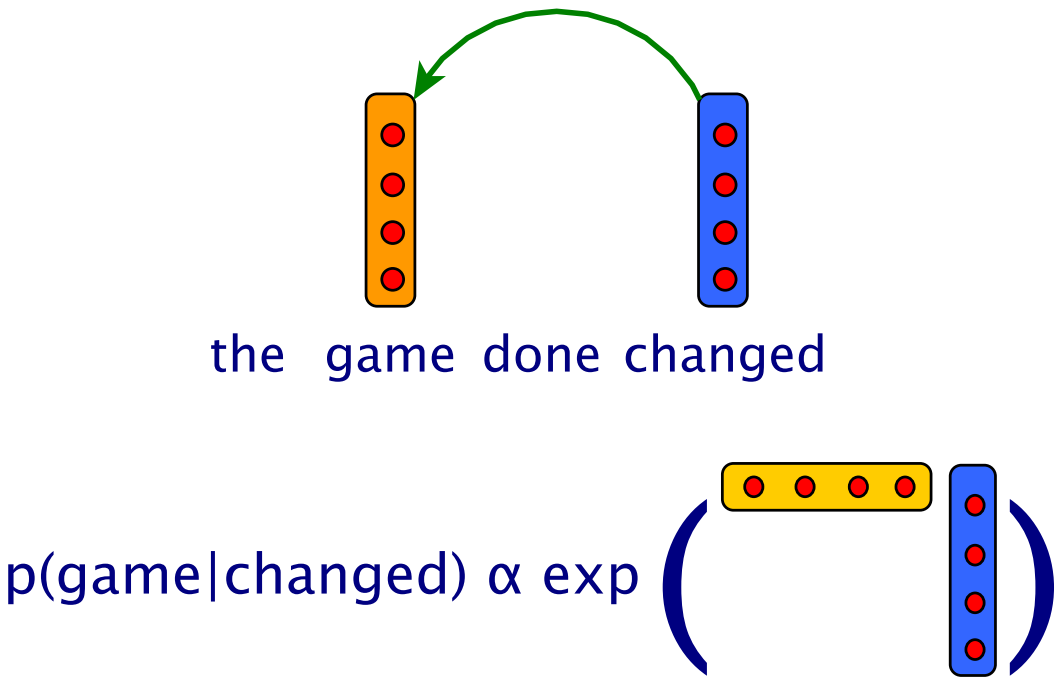
$$\Phi_1(x,y) = \begin{cases} 1, & \text{if PoS}(x)=\text{Noun} \ \& \ y=\text{in} \\ 0, & \text{otherwise} \end{cases}$$

$$\Phi_2(x,y) = \begin{cases} 1, & \text{if } x=\text{game} \ \& \ \text{PoS}(y)=\text{Prep} \\ 0, & \text{otherwise} \end{cases}$$

etc.

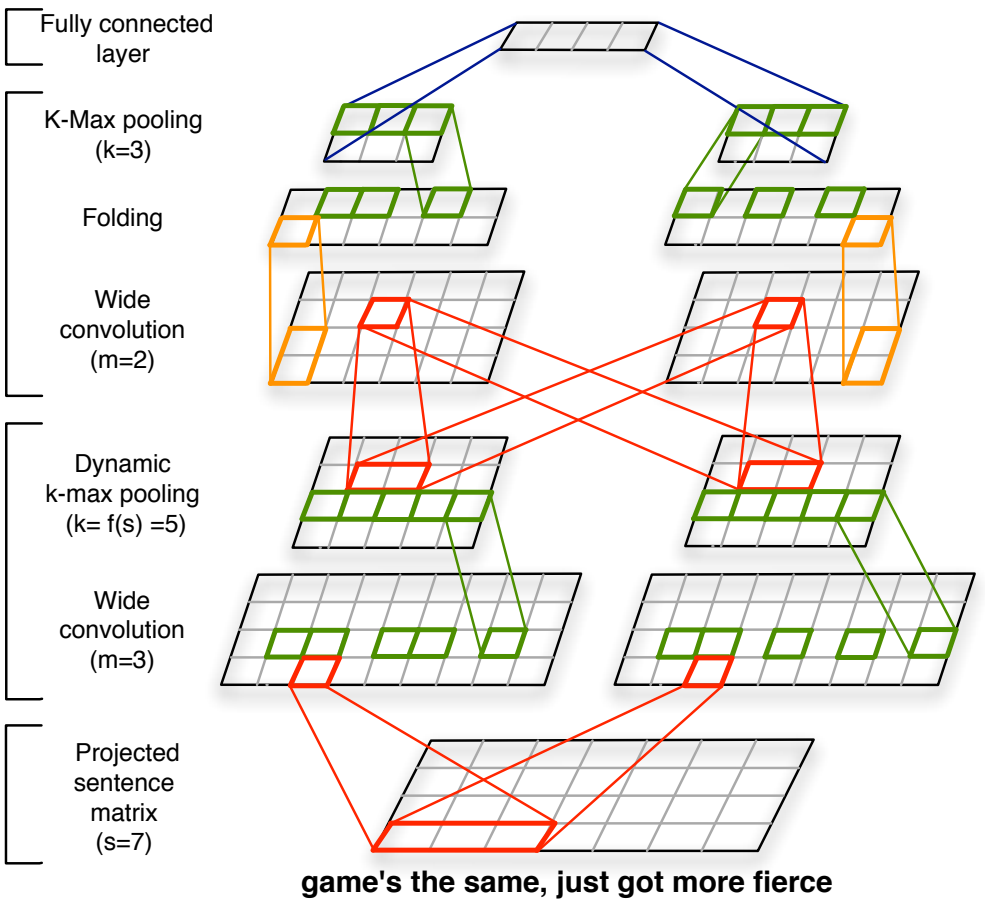
Twenty years ago log-linear models freed us from the shackles of simple multinomial parametrisations, but imposed the tyranny of feature engineering.

Features and NLP



Distributed/neural models allow us to learn shallow features for our classifiers, capturing simple correlations between inputs.

Features and NLP



Deep learning allows us to learn hierarchical generalisations. Something that is proving rather useful for vision, speech, and now NLP...

Outline

How to Represent Meaning in NLP

We can represent words using a number of approaches

- Characters
- POS tags
- Grammatical roles
- Named Entity Recognition
- Collocation and distributional representations
- Task-specific features

All of these representations can be encoded in vectors. Some of these representations capture *meaning*.

A harder problem: paraphrase detection

Q: Do two sentences (roughly) mean the same?
"He enjoys Jazz music" $\equiv$ "He likes listening to Jazz" ?

A: Use a distributional representation to find out?

A harder problem: paraphrase detection

Q: Do two sentences (roughly) mean the same?
"He enjoys Jazz music" $\equiv$ "He likes listening to Jazz" ?

A: Use a distributional representation to find out?

Most representations not sensible on the sentence level

- Characters ?
- POS tags ?
- Grammatical roles ?
- Named Entity Recognition ?
- Collocation and distributional representations ?
- Task-specific features ?

Why can't we extract hierarchical features?

The curse of dimensionality

As the dimensionality of a representation increases, learning becomes less and less viable due to sparsity.

Dimensionality for collocation

- One word per entry: Size of dictionary (small)
- One sentence per entry: Number of possible sentences (infinite)

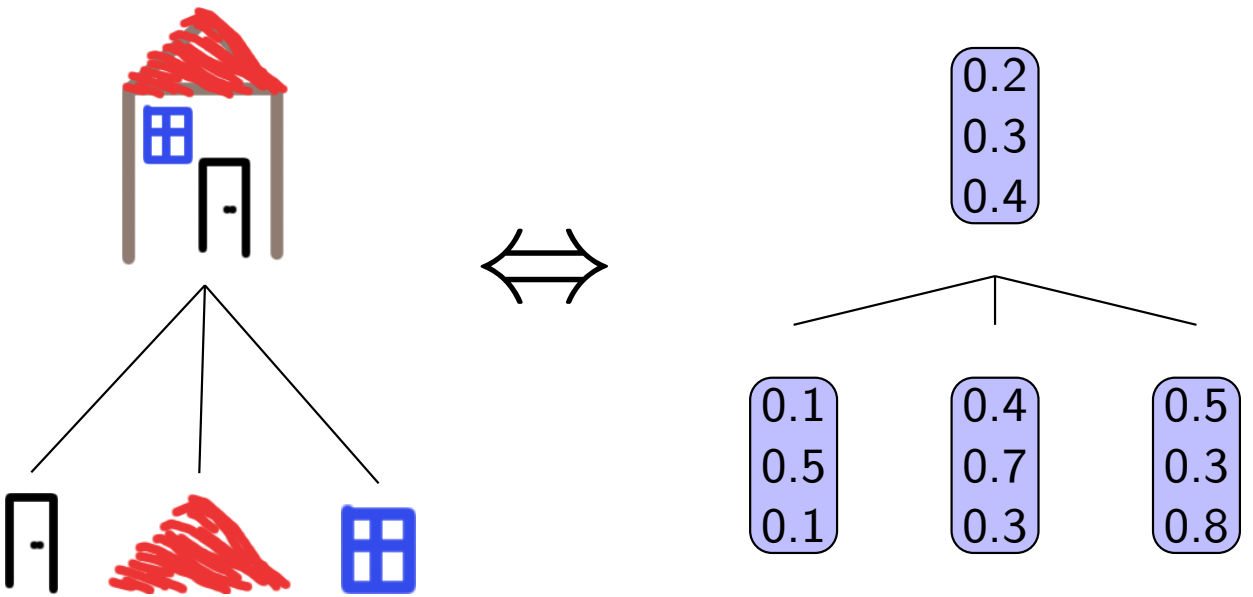
$\Rightarrow$ We need a different method for representing sentences

What is Deep Learning

Deep Learning for Language

Learning a hierarchy of features, where higher levels of abstraction are derived from lower levels.

A door, a roof, a window: It's a house



Composition

Lots of possible ways to compose vectors

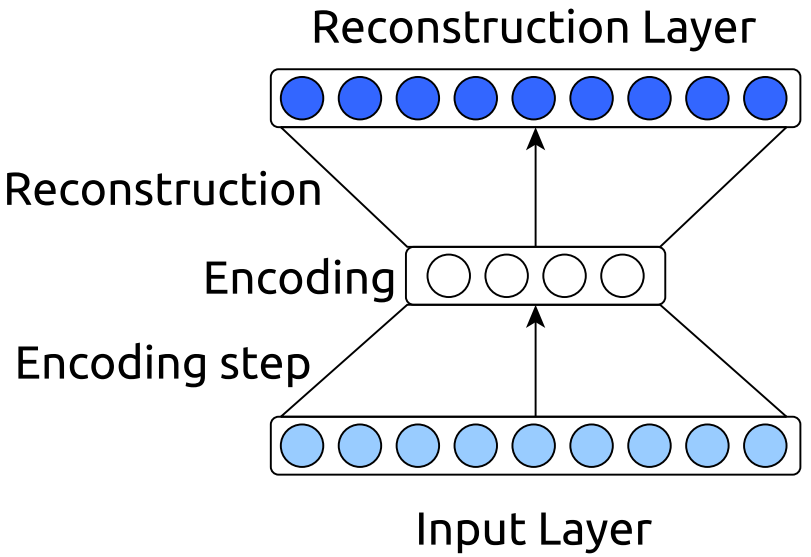
- Addition
- Multiplication
- Kronecker Product
- Tensor Magic
- Matrix-Vector multiplication
- ...

Requirements

Not commutative	Mary likes John $\neq$ John likes Mary
Encode its parts?	Magic carpet $\equiv$ Magic + Carpet
More than parts?	Memory lane $\neq$ Memory + Lane

Autoencoders

We want to ensure that the joint representation captures the meaning of its parts. We can achieve this by autoencoding our data at each step:

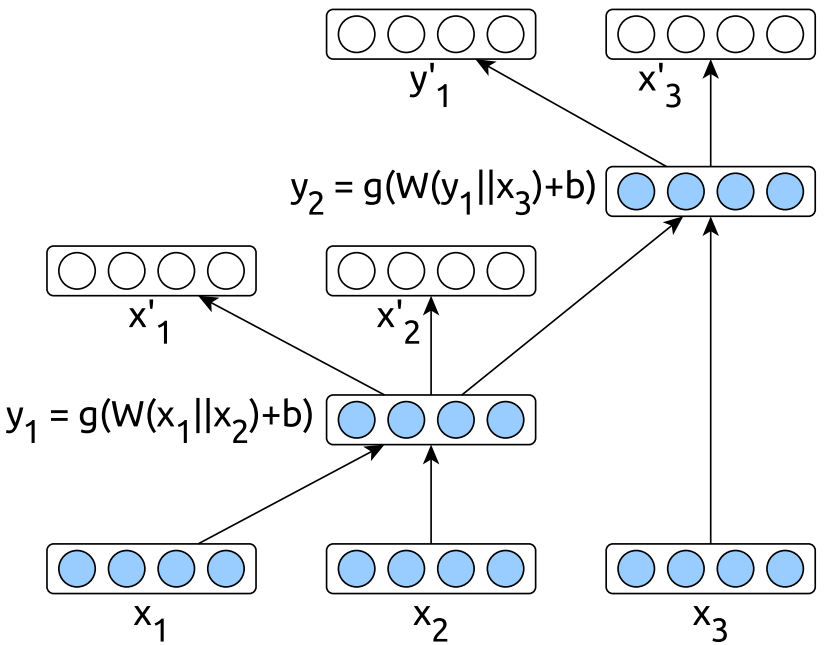


For this to work, our autoencoder minimizes an objective function over inputs $x_i, i \in N$ and their reconstructions x_i':

$$J = \frac{1}{2} \sum_{i}^{N} \left\| x_i' - x_i \right\|^2$$

Recursive Autoencoders (RAE)

We still want to learn how to represent a full sentence (or house).
To do this, we chain autoencoders to create a recursive structure.



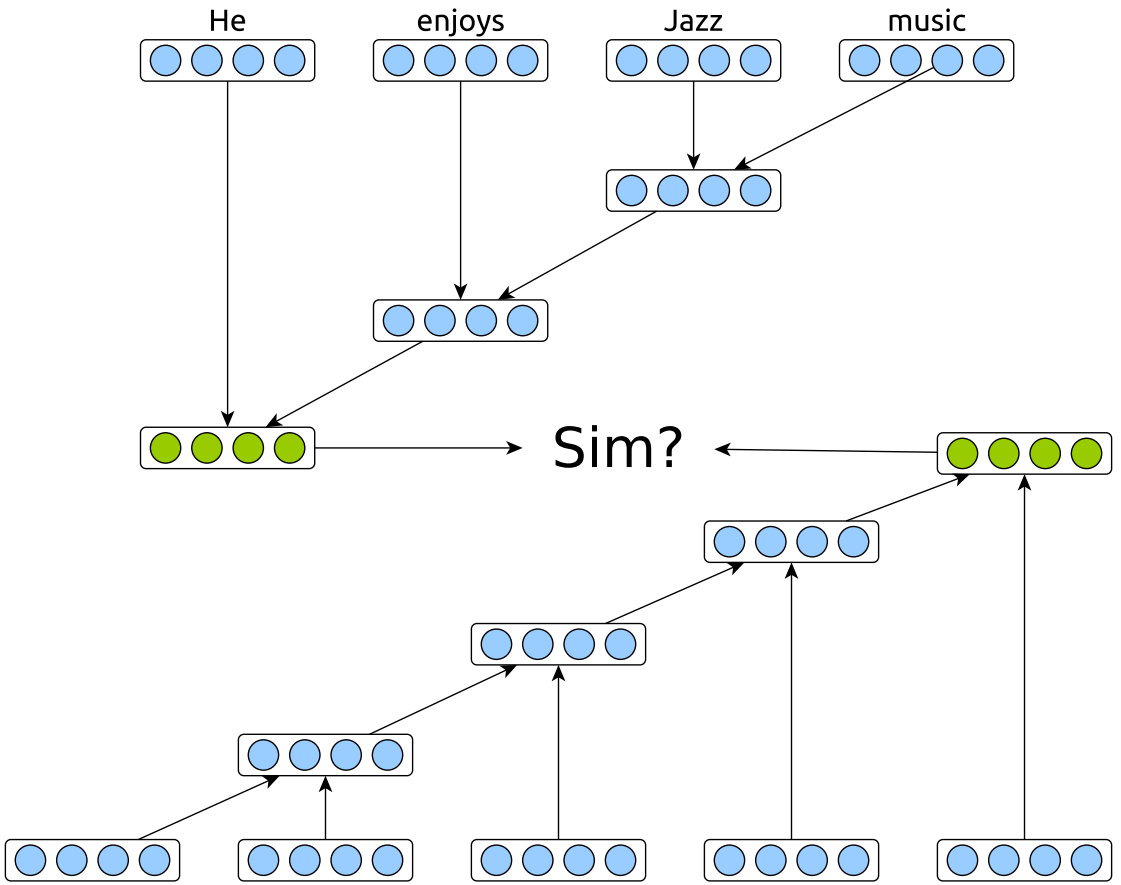
We use a composition function
$g(W * input + bias)$

g is a non-linearity (tanh, sigm)
W is a weight matrix
b is a bias

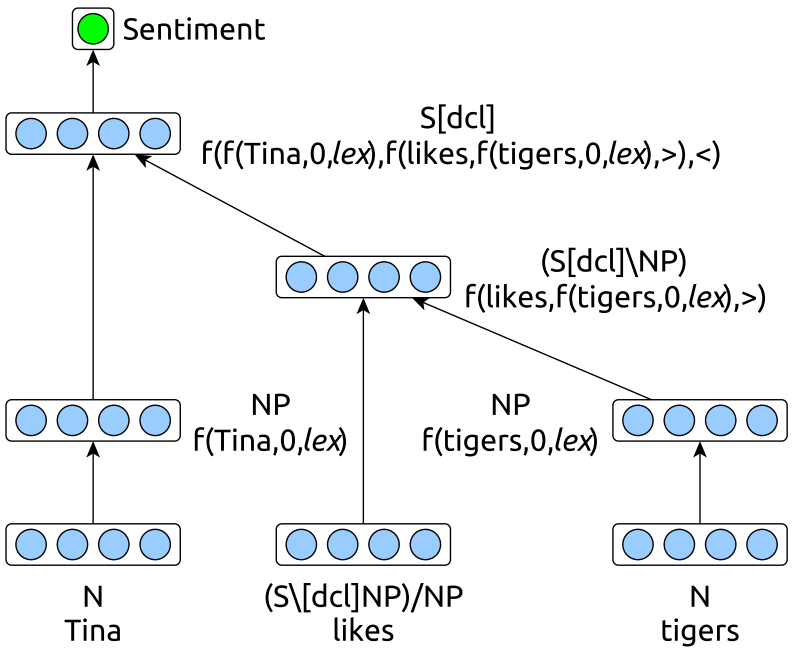
A different task: paraphrase detection

Q: Do two sentences (roughly) mean the same?
"He enjoys Jazz music" $\equiv$ "He likes listening to Jazz" ?

A: Use deep learning to find out!



Other Applications: Stick a label on top



1. Combine label and reconstruction error

$$E(N, l, \theta) =$$

$$\sum_{n \in N} E_{rec}(n, \theta) + E_{lbl}(v_n, l, \theta)$$

$$E_{rec}(n, \theta) = \frac{1}{2} \left\| [x_n \| y_n] - r_n \right\|^2$$

$$E_{lbl}(v, l, \theta) = \frac{1}{2} \| l - v \|^2$$

2. Strong results for a number of tasks:

Sentiment Analysis

Paraphrase Detection

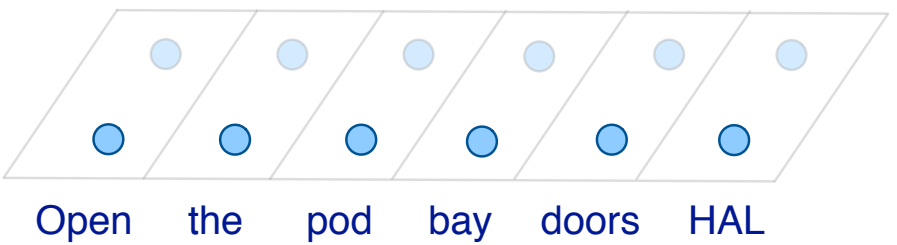
Image Search

...

Convolution Sentence Models

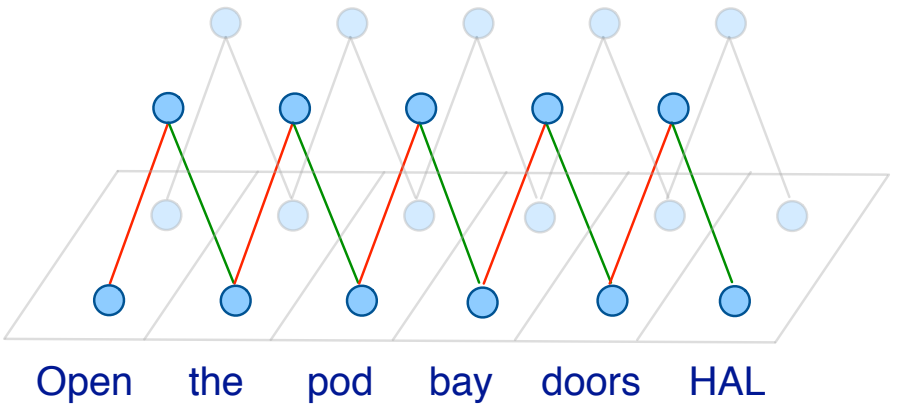
Deep learning is suppose to learn the features for us, so can we do away with all this structural engineering and forget about latent parse trees?

Convolution Sentence Models

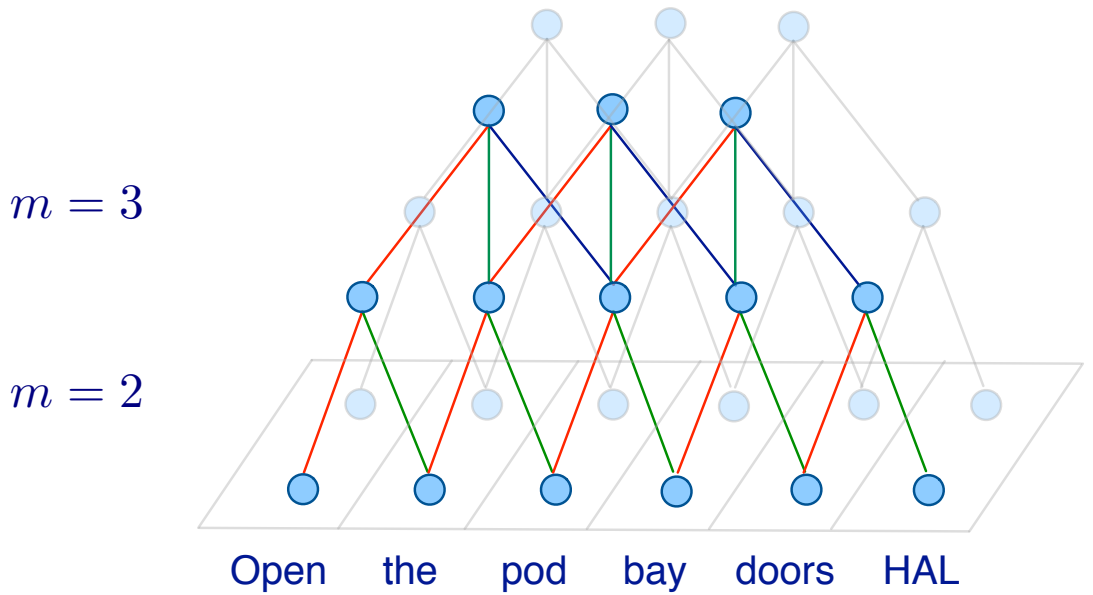


Convolution Sentence Models

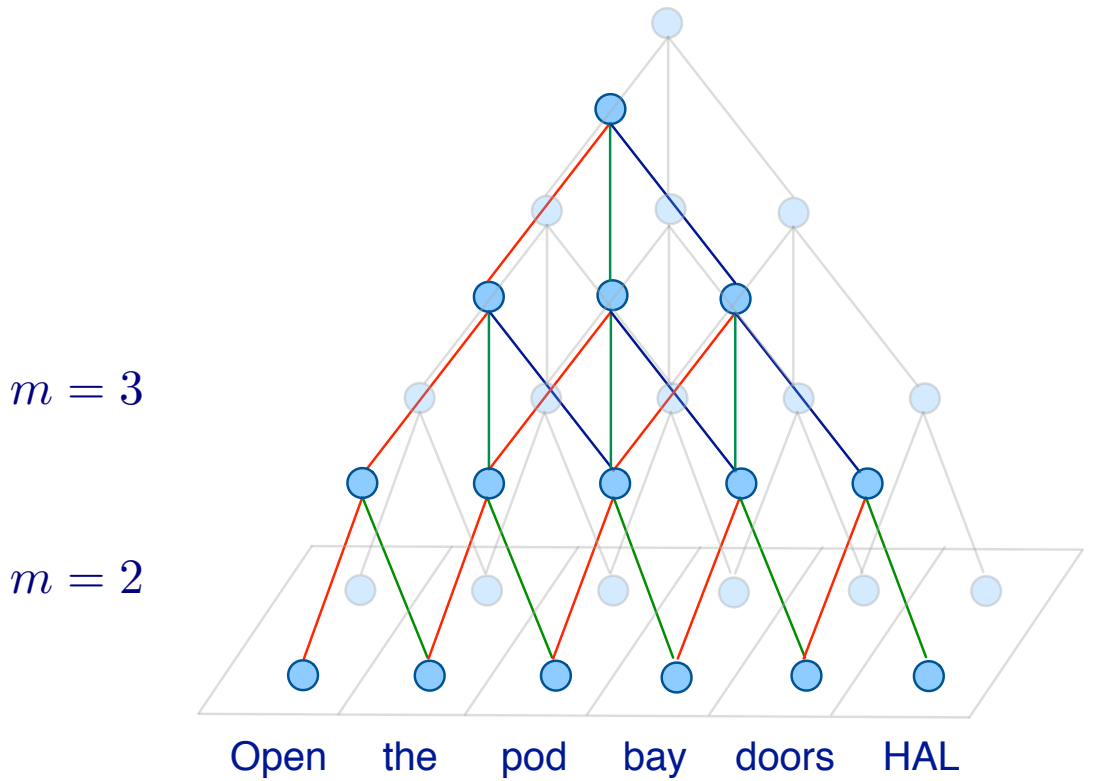
$m = 2$



Convolution Sentence Models



Convolution Sentence Models



A CSM for Dialogue Act Tagging

A: My favourite show is Masterpiece Theatre.	Statement-Non-Opinion
A: Do you like it by any chance?	Yes-No-Question
B: Oh yes!	Yes-Answers
A: You do!	Declarative Yes-No-Q
B: Yes, very much.	Yes-Answers
A: Well, wouldn't you know.	Exclamation
B: As a matter of fact, I prefer public television.	Statement-non-opinion
B: And, uh, I have, particularly enjoy English comedies.	Statement-non-opinion

A CSM for Dialogue Act Tagging

Dave: Hello HAL, do you read me HAL?

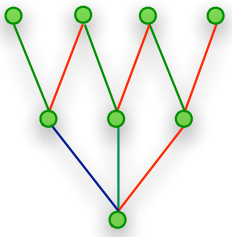
HAL: Affirmative, Dave, I read you.

Dave: Open the pod bay doors, HAL.

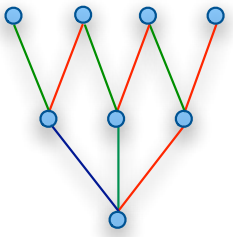
HAL: I'm sorry, Dave, I'm afraid I can't do that.

A CSM for Dialogue Act Tagging

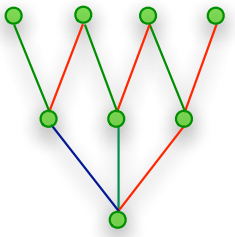
Dave: Hello HAL, do you read me HAL?



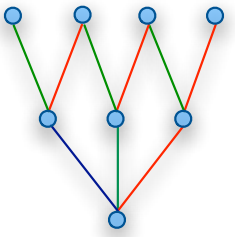
HAL: Affirmative, Dave, I read you.



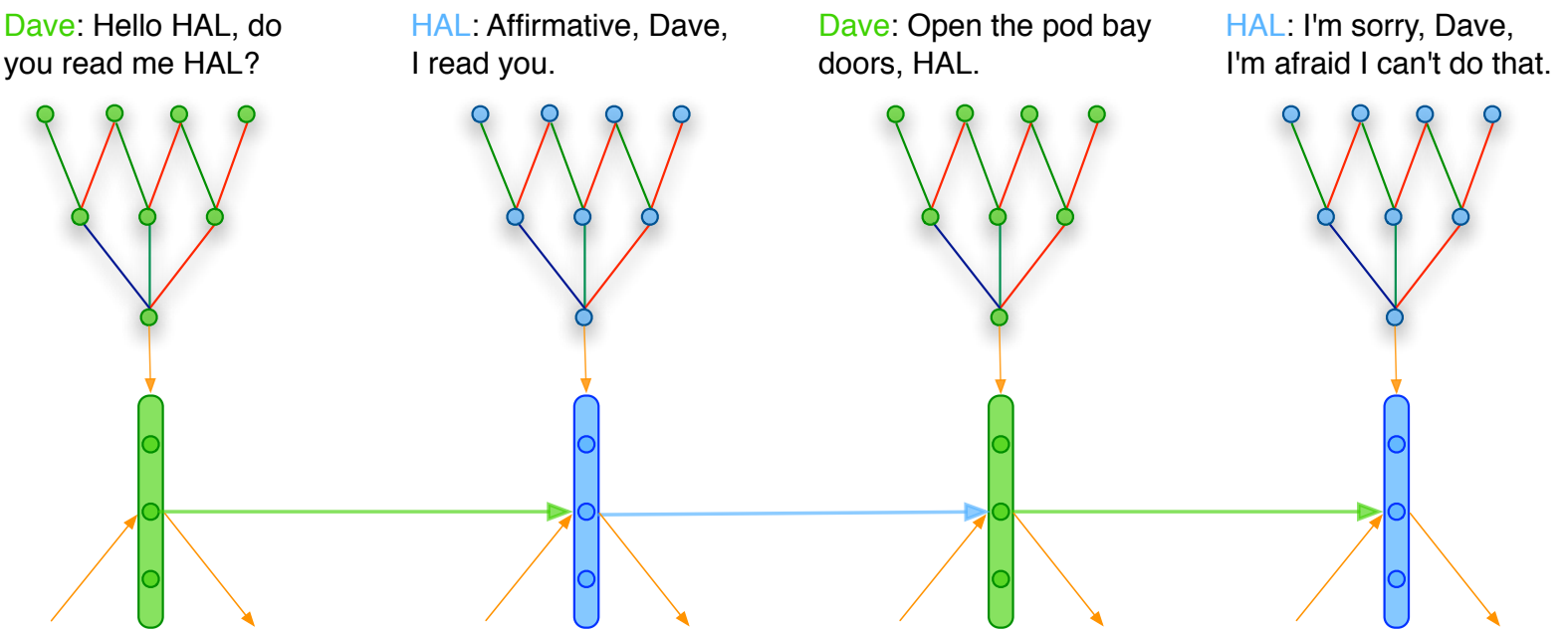
Dave: Open the pod bay doors, HAL.



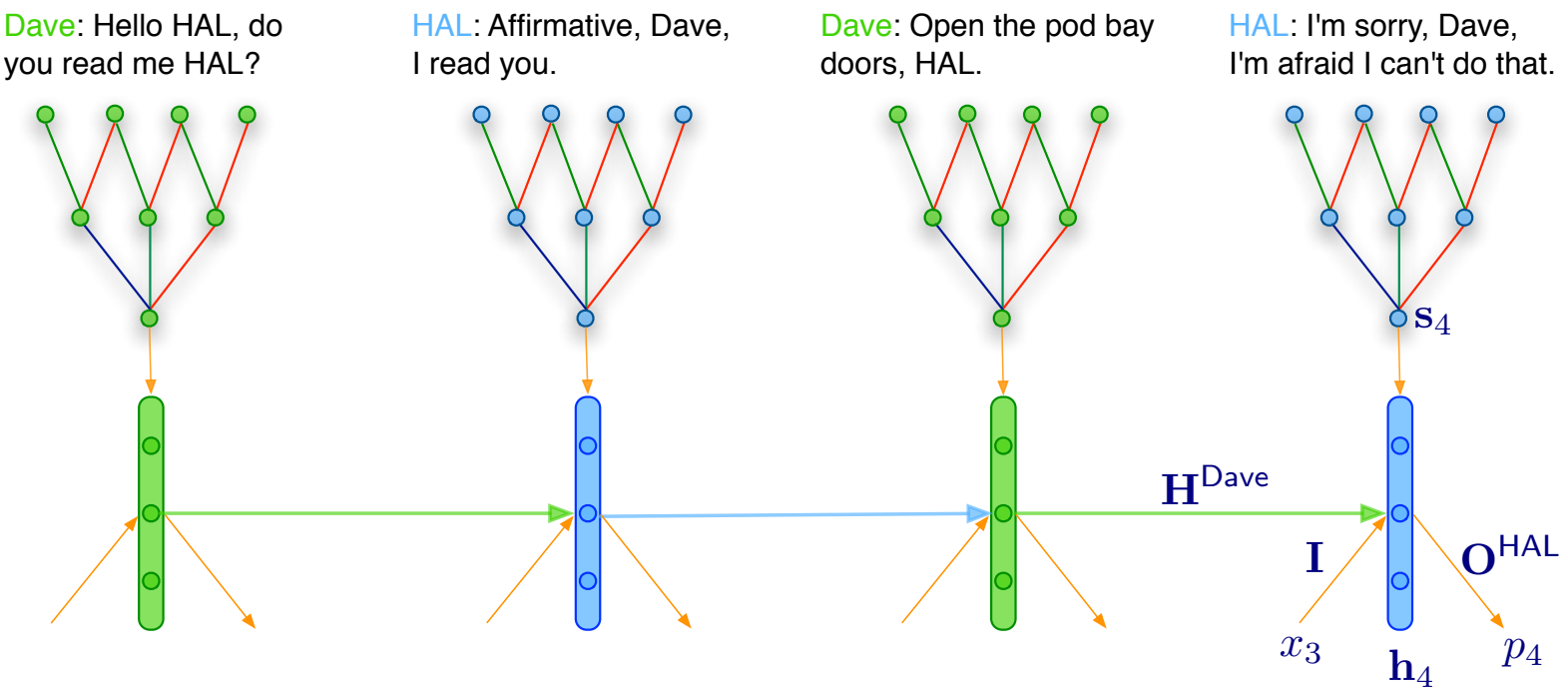
HAL: I'm sorry, Dave, I'm afraid I can't do that.



A CSM for Dialogue Act Tagging



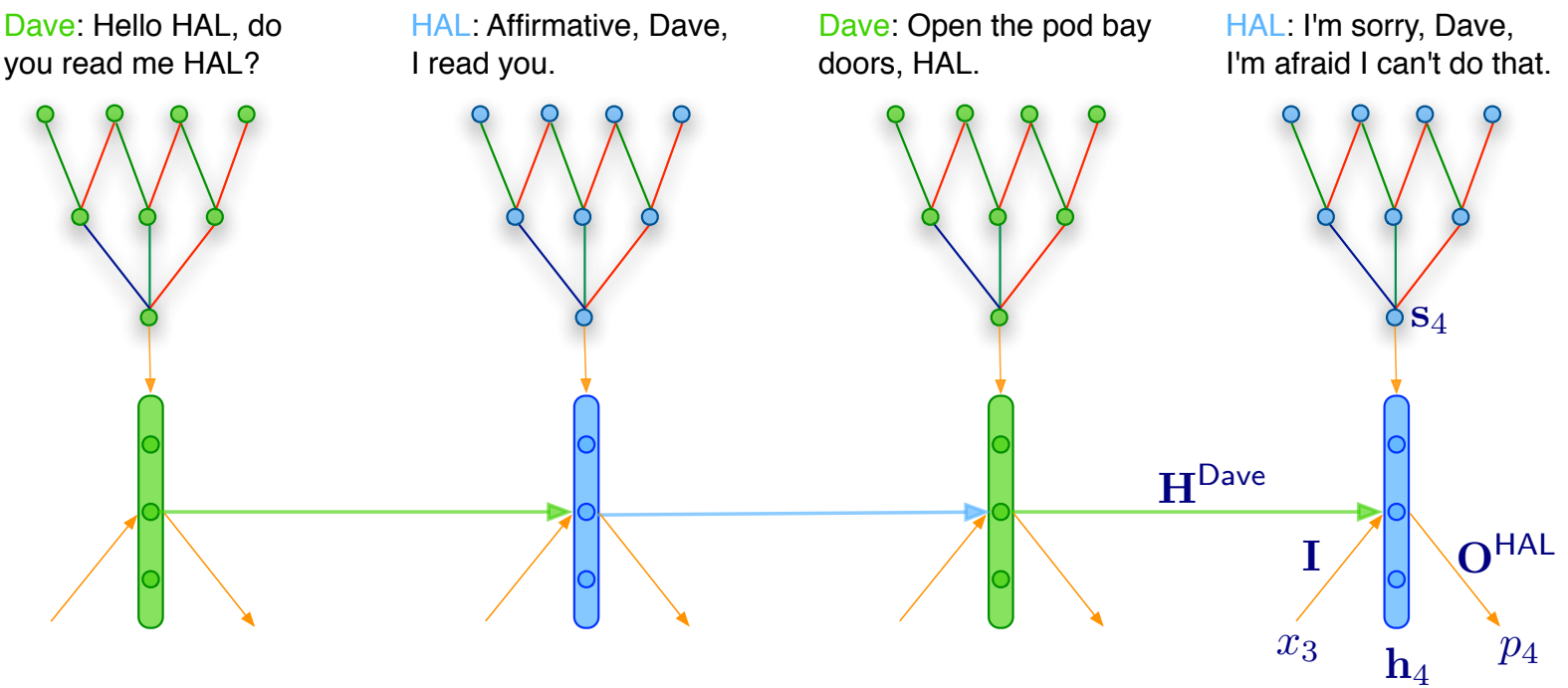
A CSM for Dialogue Act Tagging



$$\mathbf{h}_i = g(\mathbf{I}x_{i-1} + \mathbf{H}^{i-1}\mathbf{h}_{i-1} + \mathbf{S}\mathbf{s}_i)$$

$$p_i = \text{softmax}(\mathbf{O}^i\mathbf{h}_i)$$

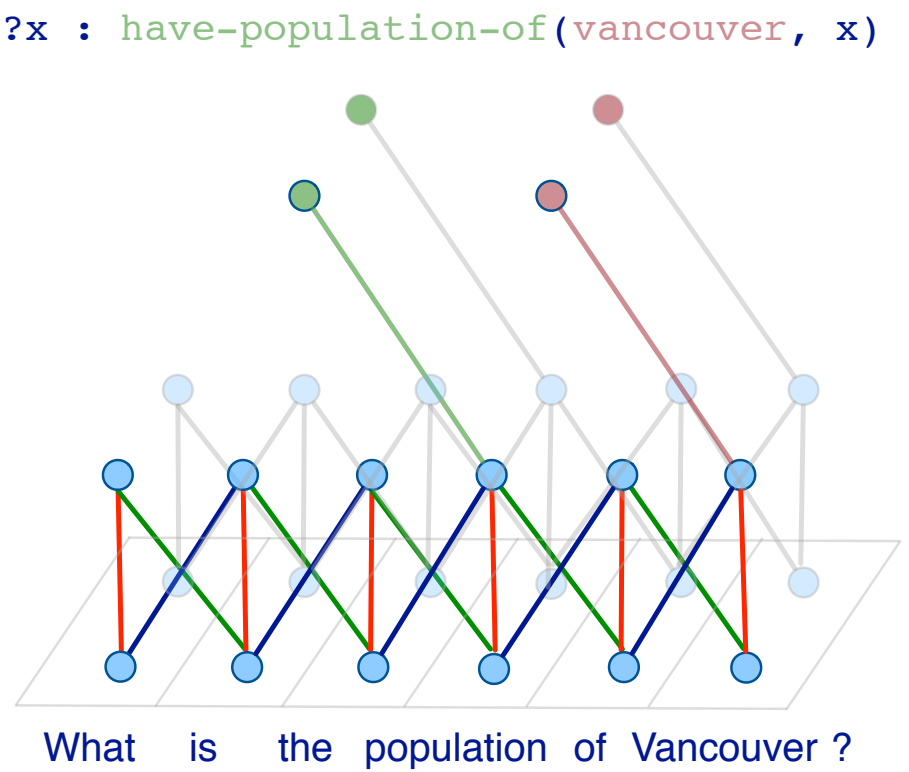
A CSM for Dialogue Act Tagging



$$\mathbf{h}_i = g(\mathbf{I}x_{i-1} + \mathbf{H}^{i-1}\mathbf{h}_{i-1} + \mathbf{S}\mathbf{s}_i)$$
$$p_i = \text{softmax}(\mathbf{O}^i\mathbf{h}_i)$$

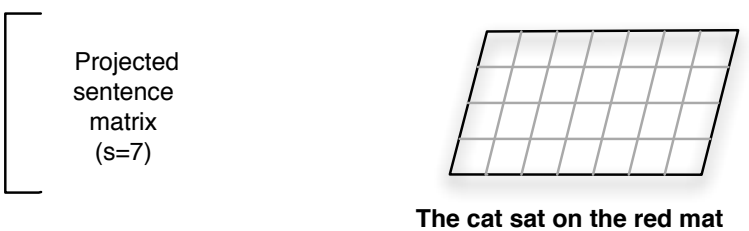
State of the art results while allowing online processing of dialogue.

Convolution Sentence Models: Question Answering

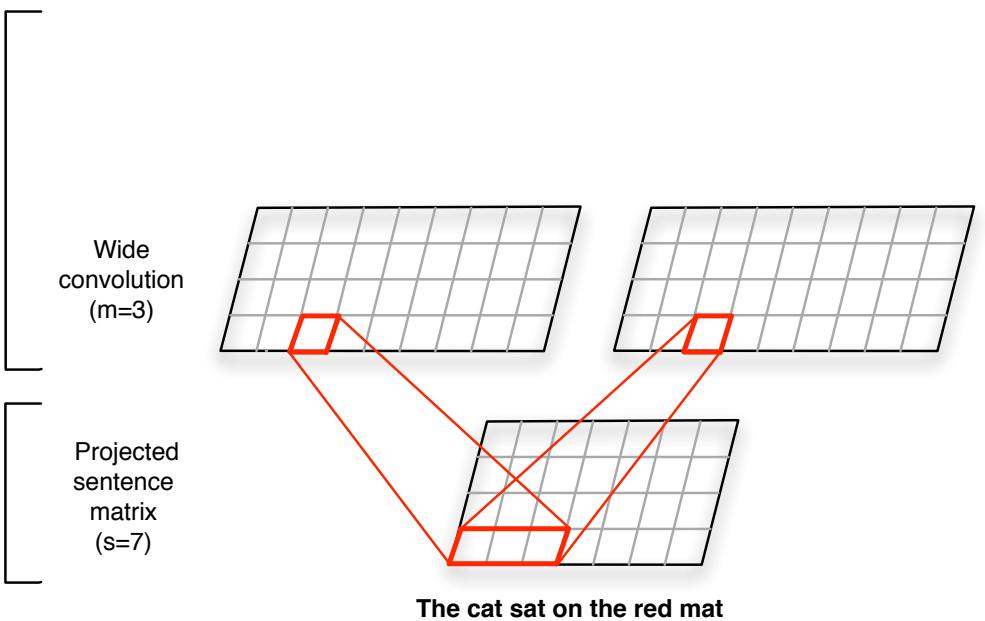


Competitive with a template based approach with lots of hand engineered features.

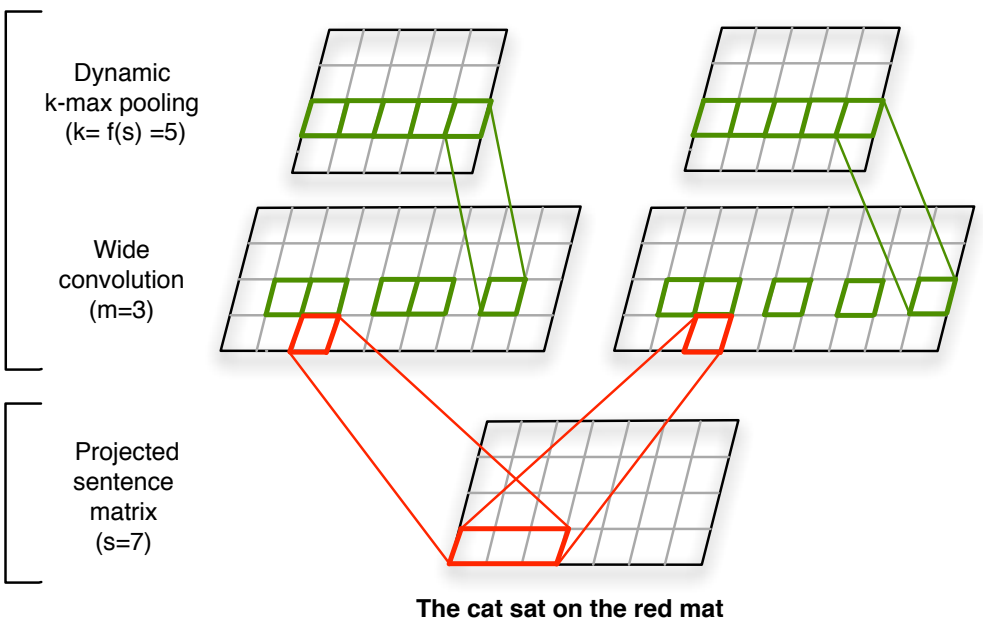
Convolution Sentence Models



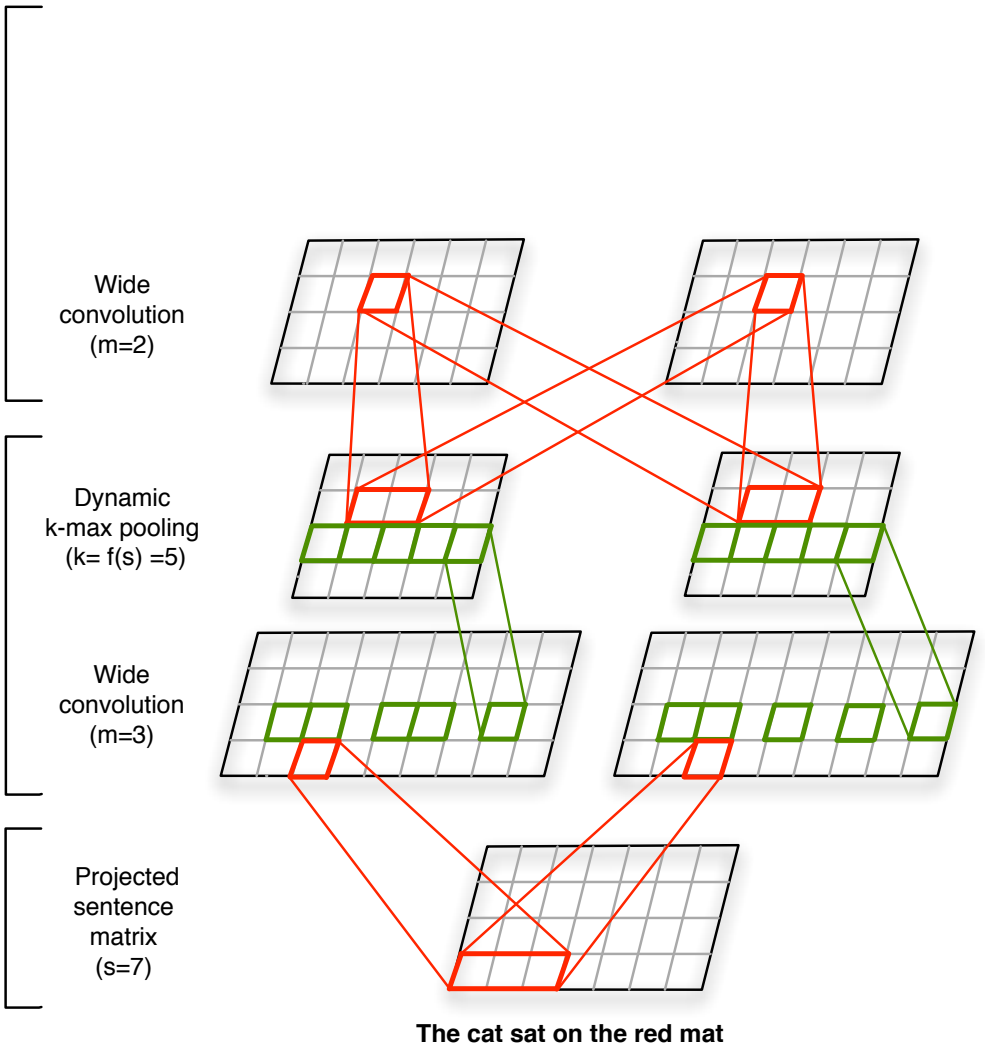
Convolution Sentence Models



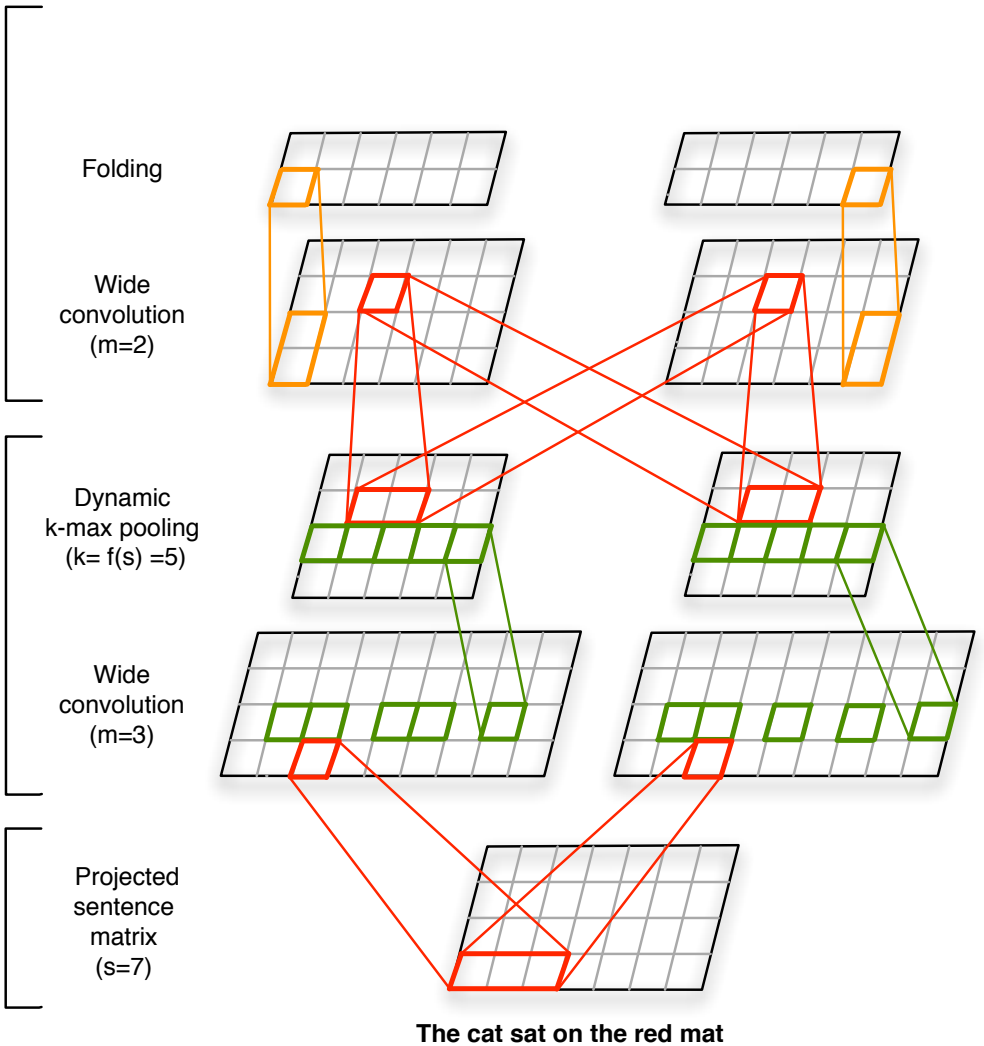
Convolution Sentence Models



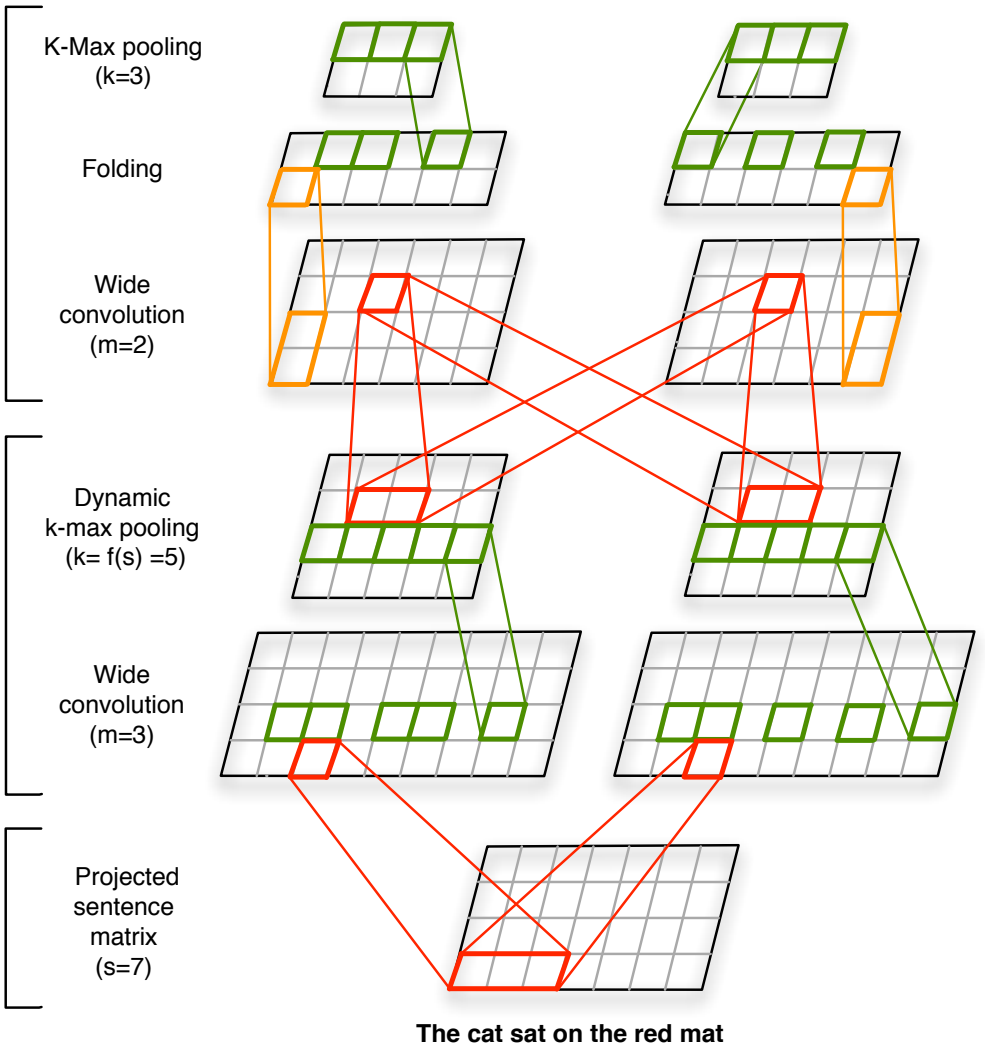
Convolution Sentence Models



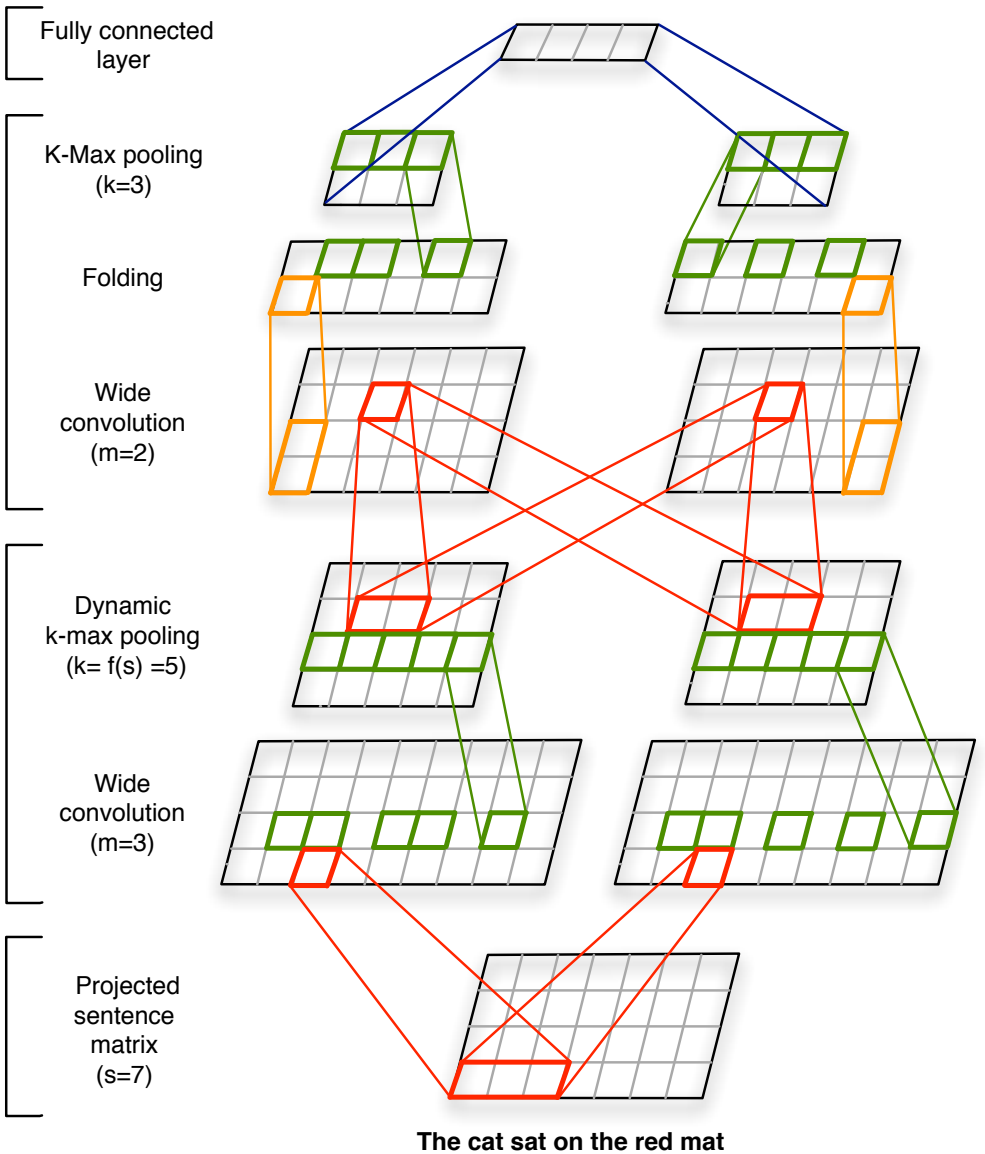
Convolution Sentence Models



Convolution Sentence Models



Convolution Sentence Models



Small Sentiment Task

	Five-class (%)	Binary (%)
NB	41.0	81.8
SVM	40.7	79.4
BINB	41.9	83.1
RECNTN	45.7	85.4
MAX-TDNN	37.4	77.1
NBOW	42.4	80.5
DCNN	**48.5**	**86.8**

Sentiment prediction on the Stanford movie reviews dataset.

Large Sentiment Task

	Accuracy (%)
SVM	81.6
NB	82.7
MAXENT	83.0
MAX-TDNN	78.8
NBOW	80.9
DCNN	**87.4**

Accuracy on the larger Twitter sentiment dataset.

Question Classification Task

Classifier	Features	Acc. (%)
HIER	unigram, POS, head chunks NE, semantic relations	91.0
MAXENT	unigram, bigram, trigram POS, chunks, NE, supertags CCG parser, WordNet	92.6
MAXENT	unigram, bigram, trigram POS, wh-word, head word word shape, parser hypernyms, WordNet	93.6
SVM	unigram, POS, wh-word head word, parser hypernyms, WordNet 60 hand-coded rules	95.0
MAX-TDNN	unsupervised vectors	84.4
NBOW	unsupervised vectors	88.2
DCNN	unsupervised vectors	93.0

Six-way question classification on the TREC questions dataset, e.g.

Input: How far is it from Denver to Aspen ?

Output: NUMBER

Feature: not only . . . but also

not	only	manufactured	,	but	also	so
while	not	all	transitions	to	are	so
s	not	there	yet			but
not	all	transitions	to	are	so	,
may	not	be	new	,	but	australian
feels	not	only	manufactured	,	but	also
s	not	merely	unwatchable	,	but	also
land	than	crash	,	but	ultimately	serving
least	surprising	,	it	is	still	ultimately
great	bond	movie	,	but	it	is

Feature: as . . . as . . . as

as	predictable	and	as	lowbrow	as	the
as	lively	and	as	fun	as	it
,	confusing	spectacle	,	one	that	may
that	hinges	on	its	casting	,	and
cinematic	high	crime	,	one	that	brings
as	an	athlete	as	well	as	an
its	audience	and	its	source	material	.
,	and	lane	as	vincent	,	the
as	lo	fi	as	the	special	effects
age	story	restraint	as	well	as	warmth

Feature: positivity

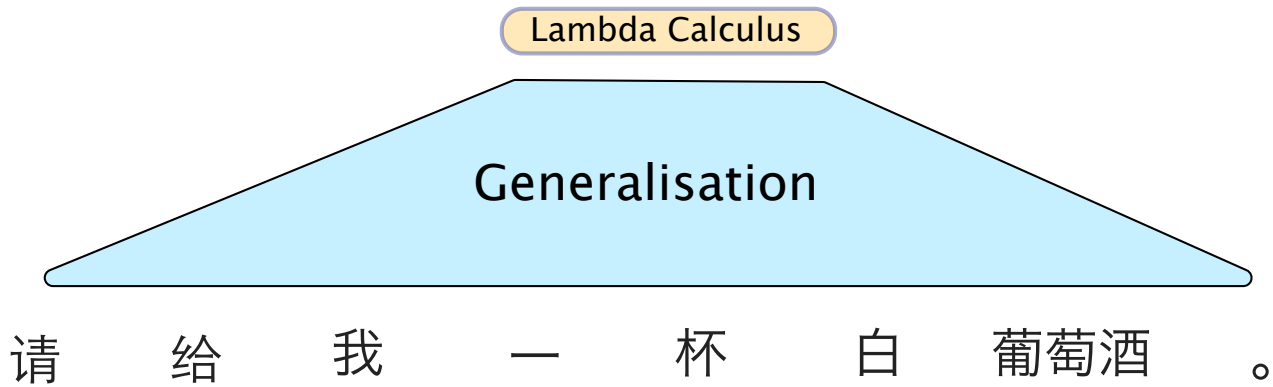
startling	film	that	gives	you	a	fascinating
well	written	,	nicely	acted	and	beautifully
best	from	his	large	cast	in	beautifully
strong	,	credible	performances	from	the	whole
compelling	journey	...	and	``	his	best
be	a	joyful	or	at	least	fascinating
throughout	is	daring	,	inventive	and	impressive
enjoyable	film	for	the	family	,	amusing
originality	it	makes	up	for	in	intelligence
charming	,	quirky	and	paced	scottish	comedy

Outline

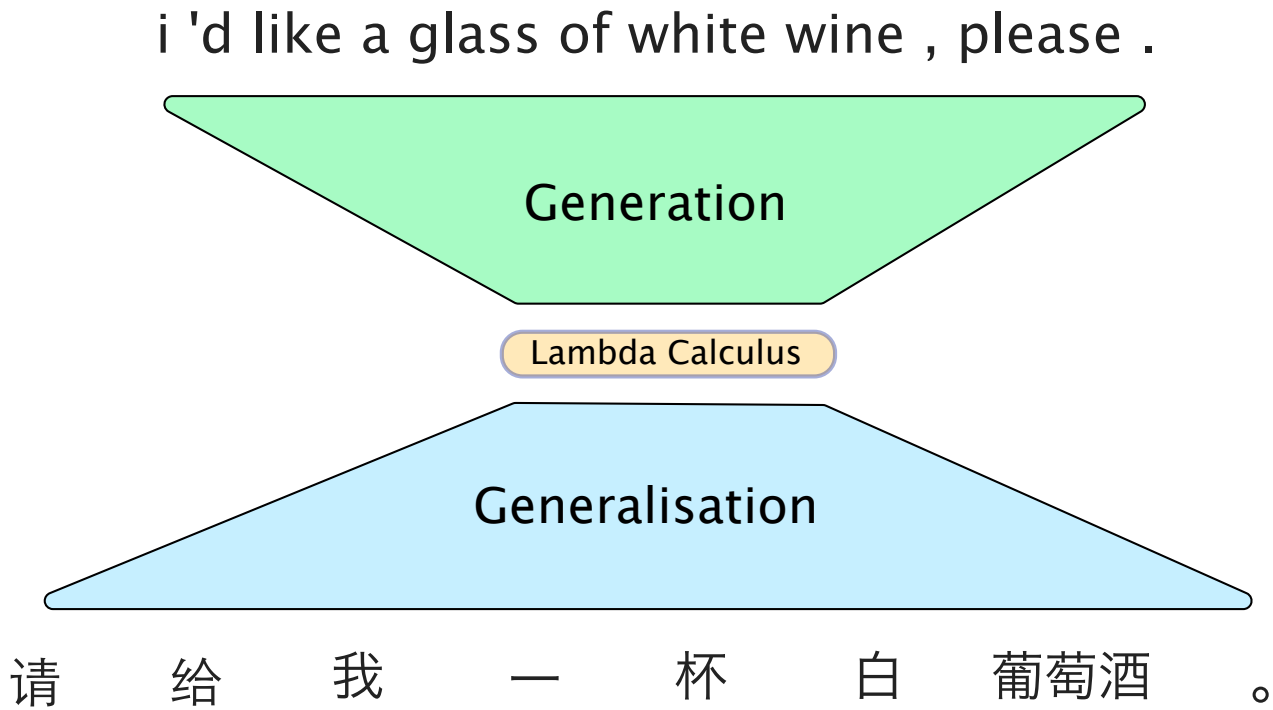
Generalisation in MT

请 给 我 一 杯 白 葡萄酒 。

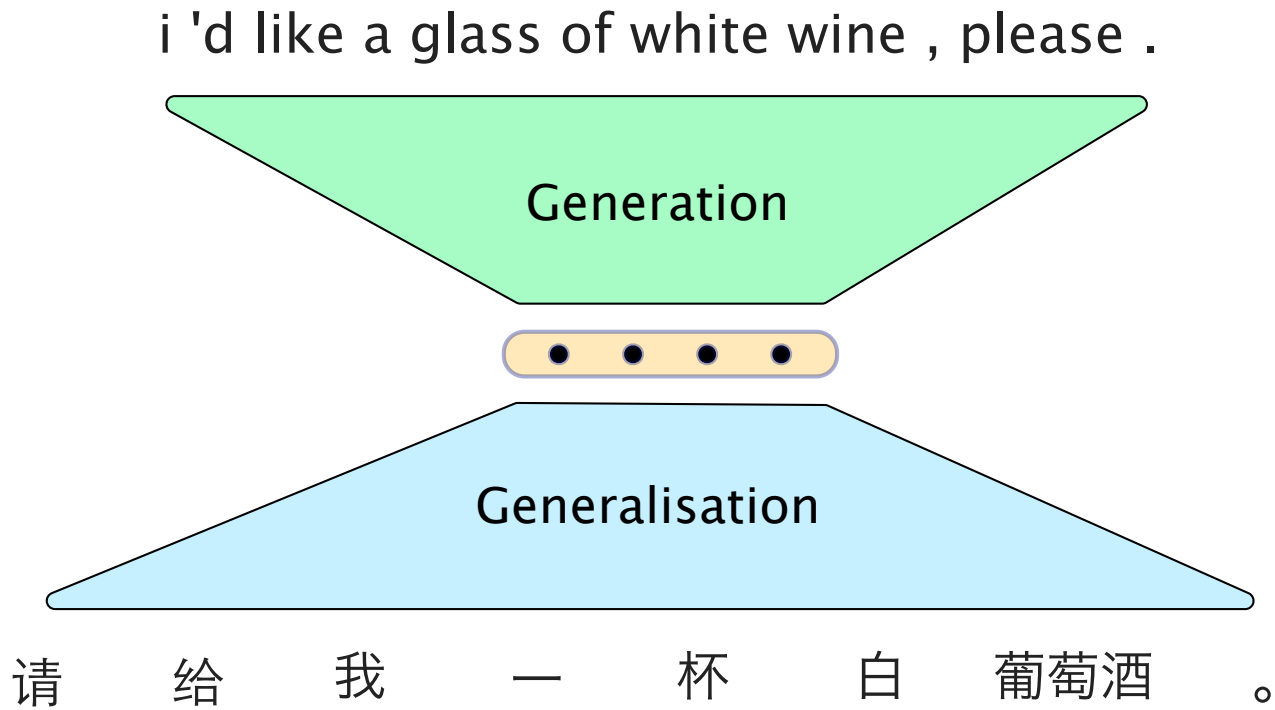
Generalisation in MT



Generalisation in MT



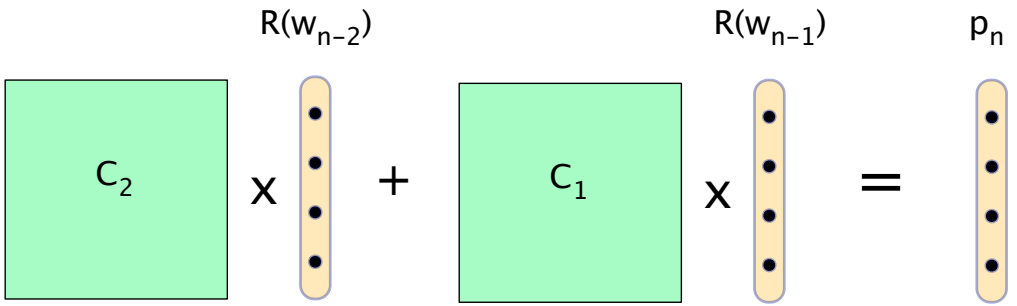
Generalisation in MT



Formal logical representations are very hard to learn from data. Let us optimistically assume a vector space and see how we go.

Generation

A simple distributed representation language model:

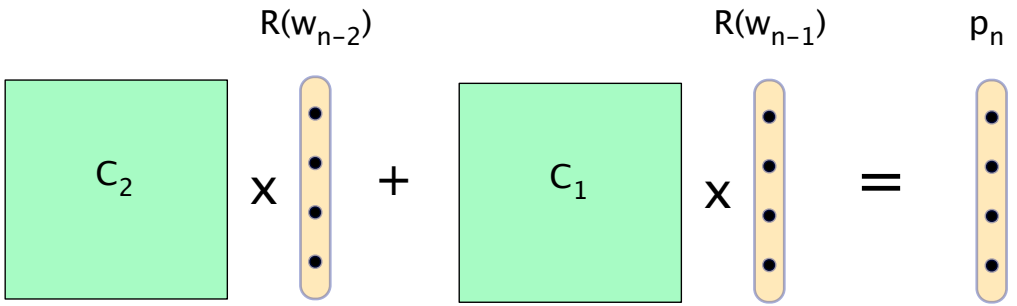


$$p_n = C_{n-2}R(w_{n-2}) + C_{n-1}R(w_{n-1})$$
$$p(w_n|w_{n-1}, w_{n-2}) \propto \exp\left(R(w_n)^T p_n\right)$$

This is referred to as a *log-bilinear model.*

Generation

A simple distributed representation language model:

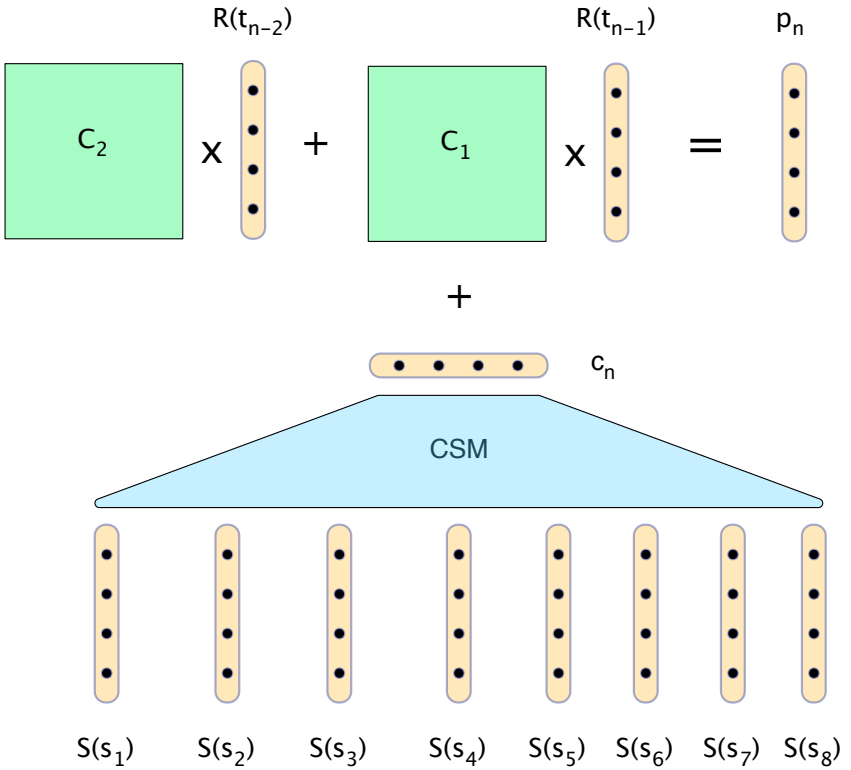


$$p_n = C_{n-2}R(w_{n-2}) + C_{n-1}R(w_{n-1})$$

$$p(w_n|w_{n-1}, w_{n-2}) \propto \exp\left(R(w_n)^T \sigma(p_n)\right)$$

Adding a non-linearity gives a version of what is often called a neural, or continuous space, LM.

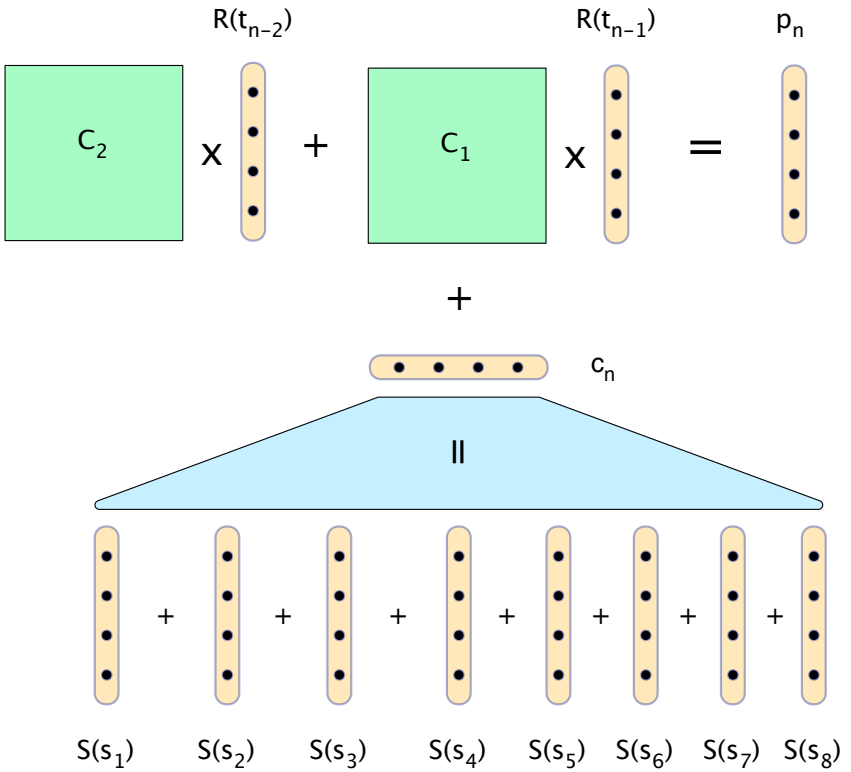
Conditional Generation



$$p_n = C_{n-2}R(t_{n-2}) + C_{n-1}R(t_{n-1}) + \texttt{CSM}(n, \mathbf{s})$$

$$p(t_n | t_{n-1}, t_{n-2}, \mathbf{s}) \propto \exp\left(R(t_n)^T \sigma(p_n)\right)$$

Conditional Generation: A Naive First Model



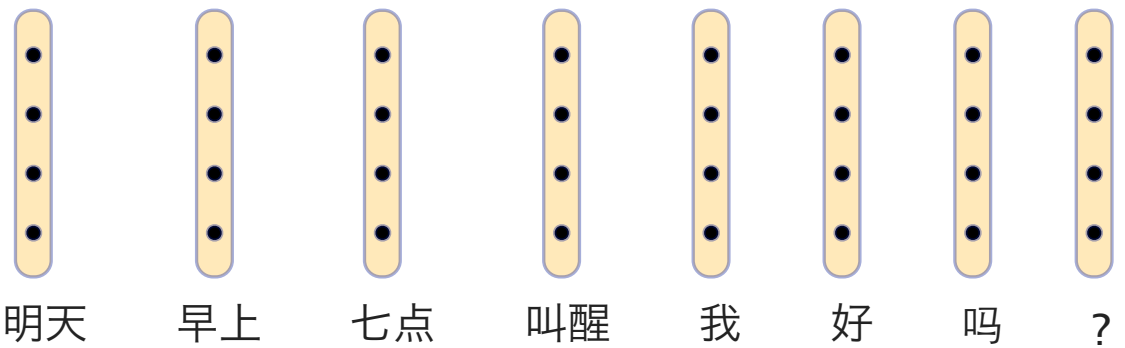
$$p_n = C_2 R(t_{n-2}) + C_1 R(t_{n-1}) + \sum_{j=1}^{|\mathbf{s}|} S(s_j)$$

$$p(t_n | t_{n-1}, t_{n-2}, \mathbf{s}) \propto \exp\left(R(t_n)^T \sigma(p_n)\right)$$

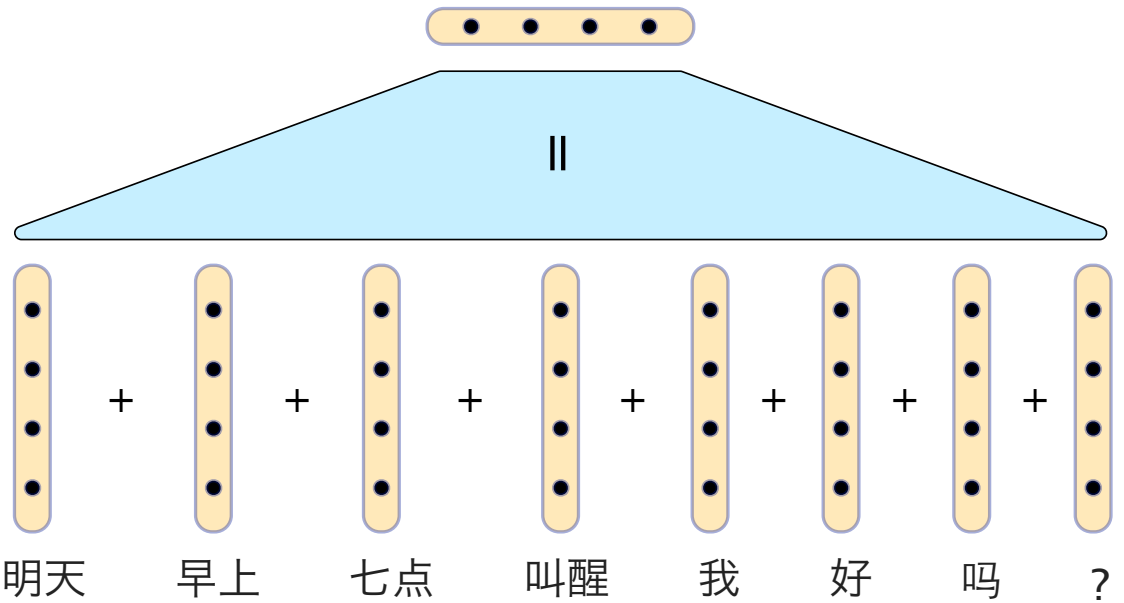
Conditional Generation: A Naive First Model

明天 早上 七点 叫醒 我 好 吗 ?

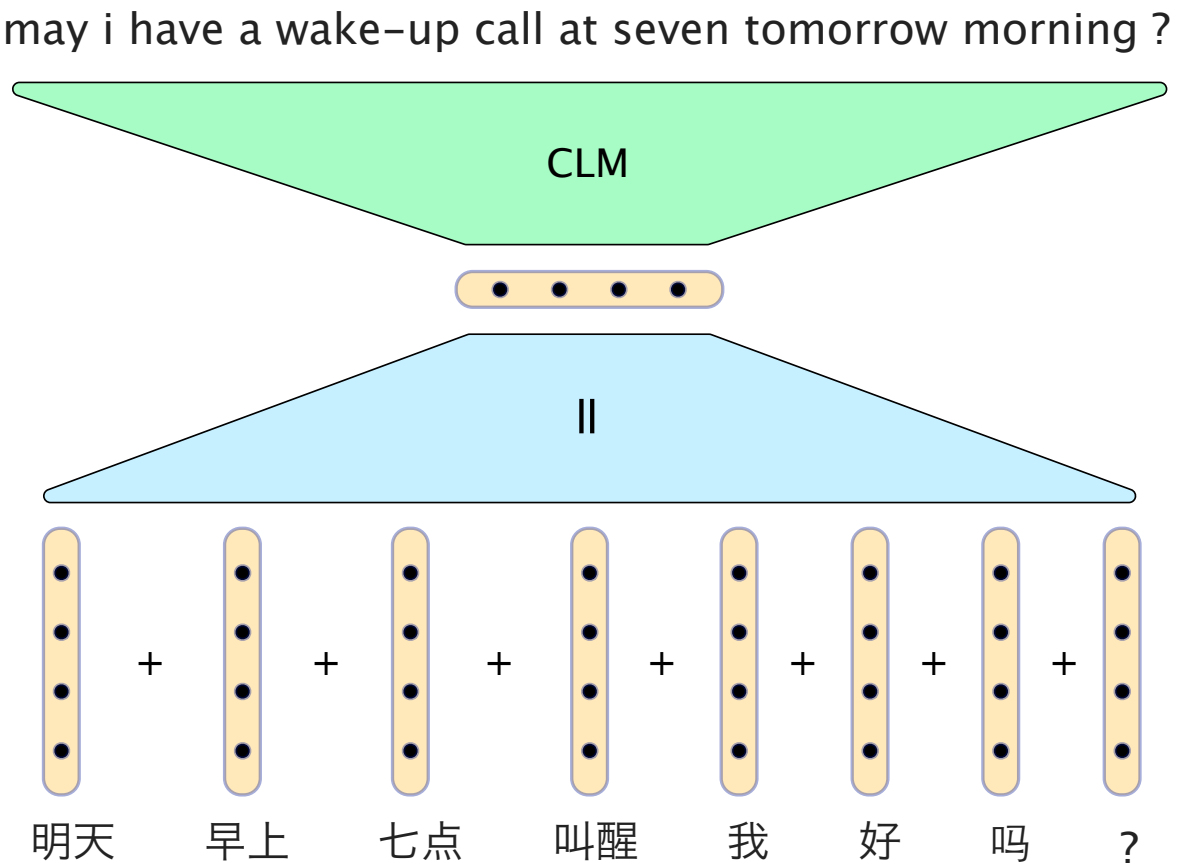
Conditional Generation: A Naive First Model



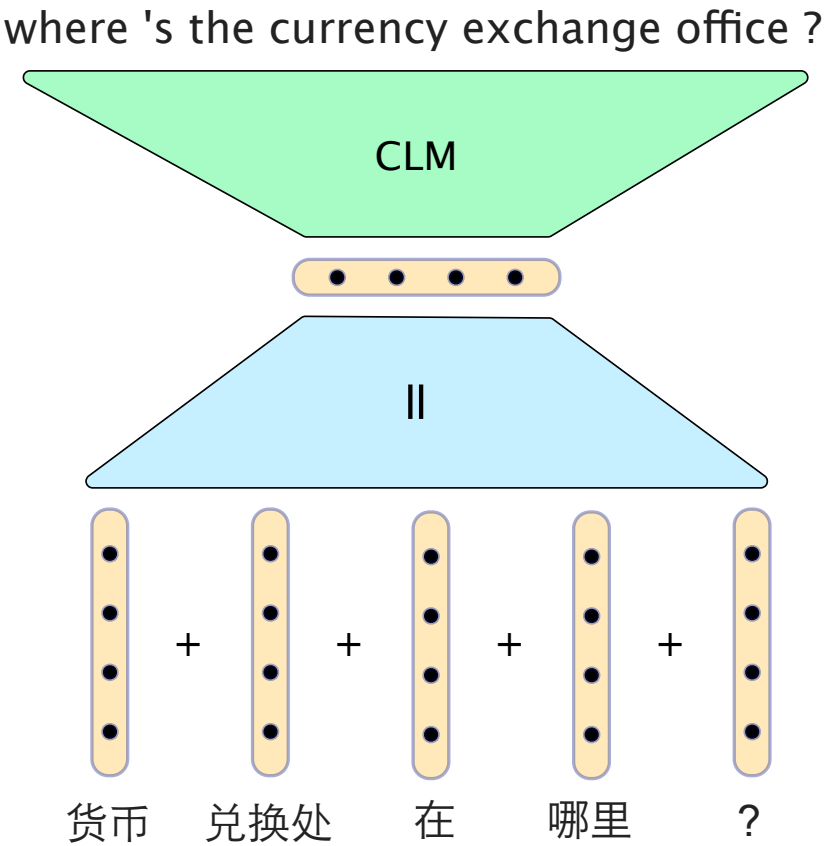
Conditional Generation: A Naive First Model



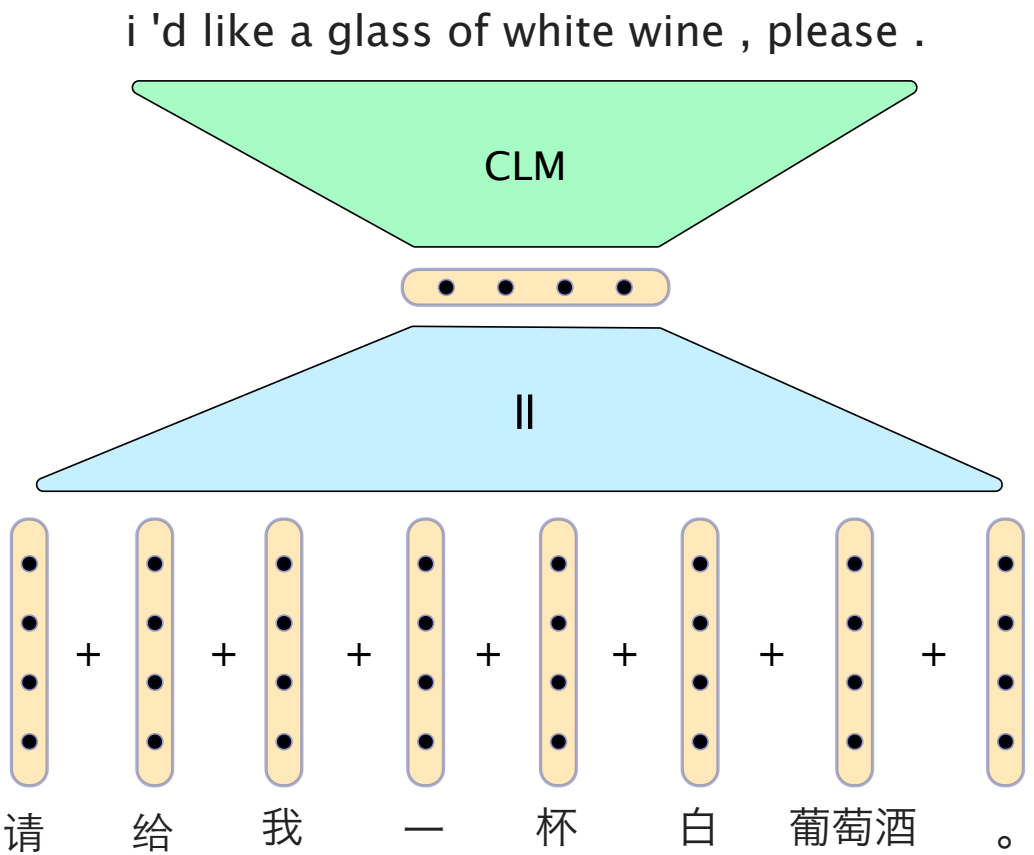
Conditional Generation: A Naive First Model



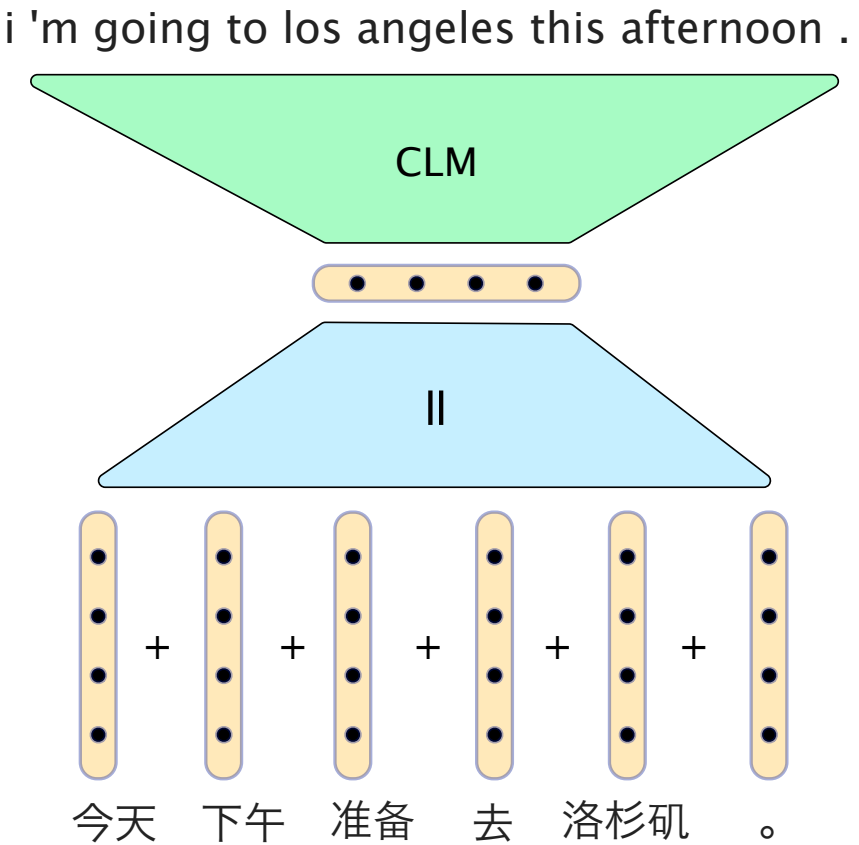
Conditional Generation: A Naive First Model



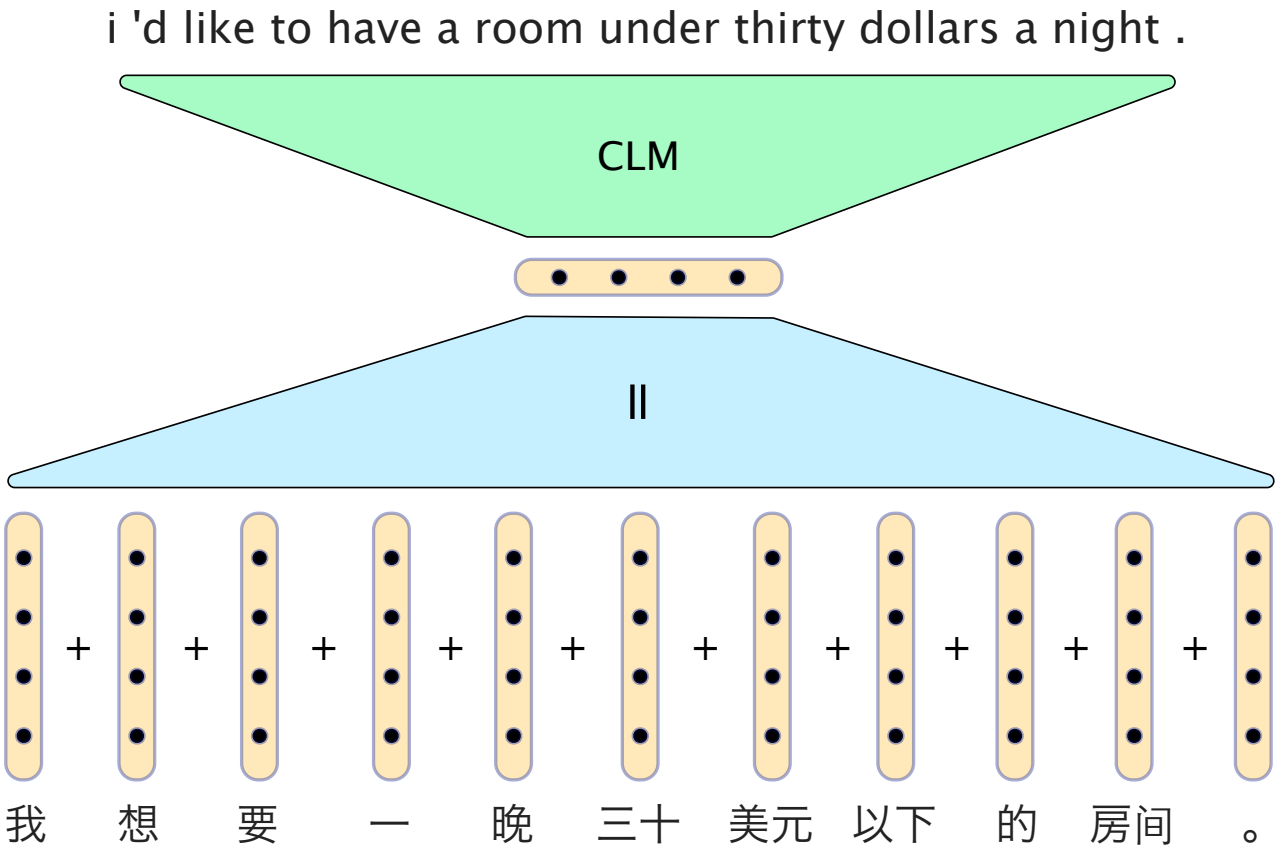
Conditional Generation: A Naive First Model



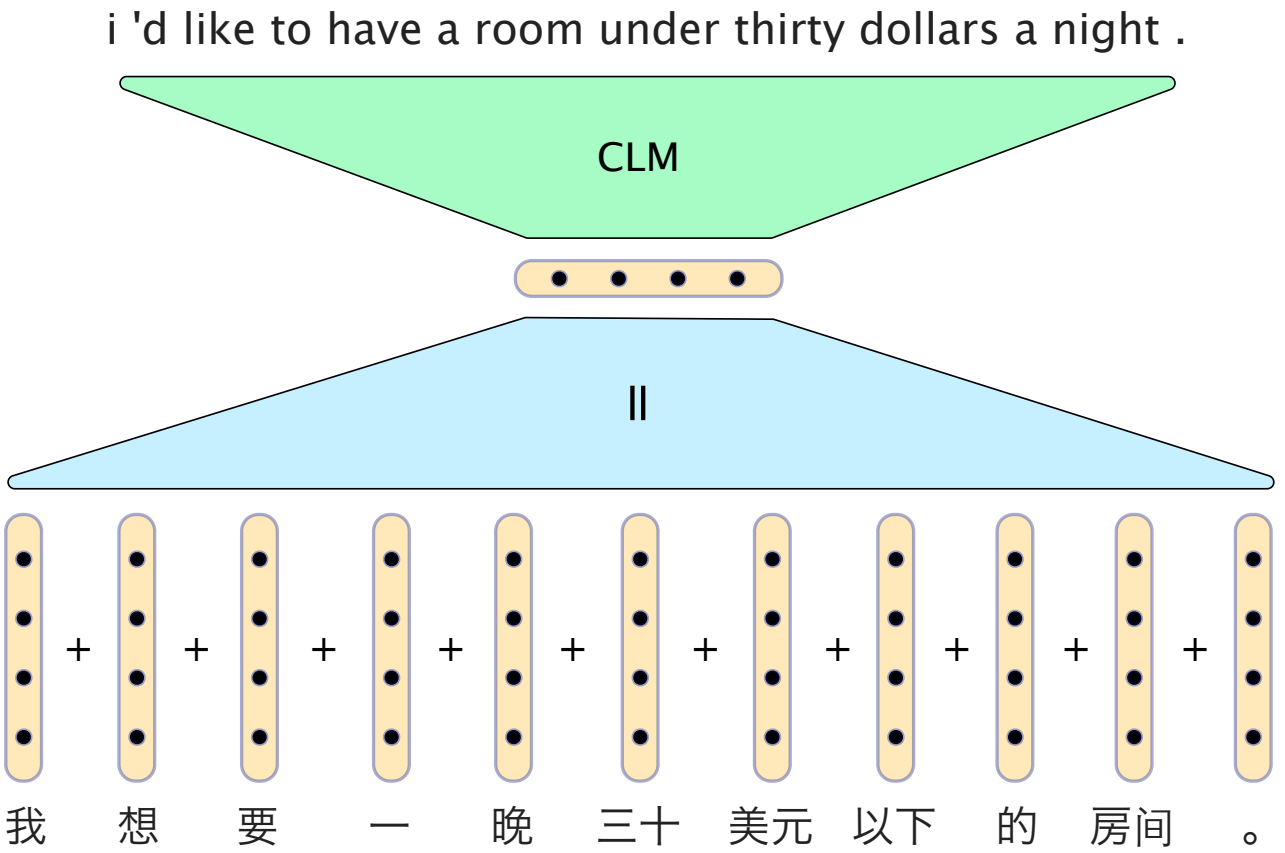
Conditional Generation: A Naive First Model



Conditional Generation: A Naive First Model



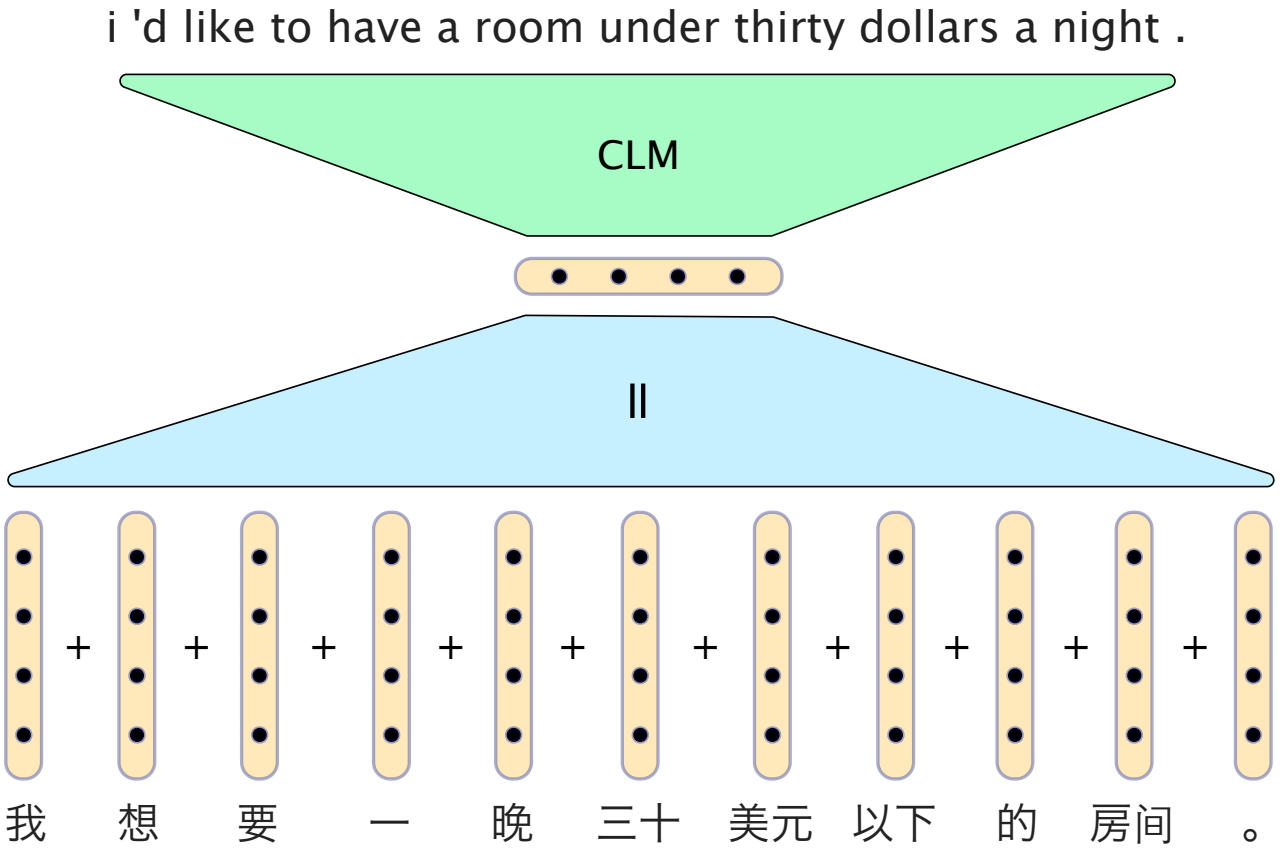
Conditional Generation: A Naive First Model



Rough Gloss

I would like a night thirty dollars under room.

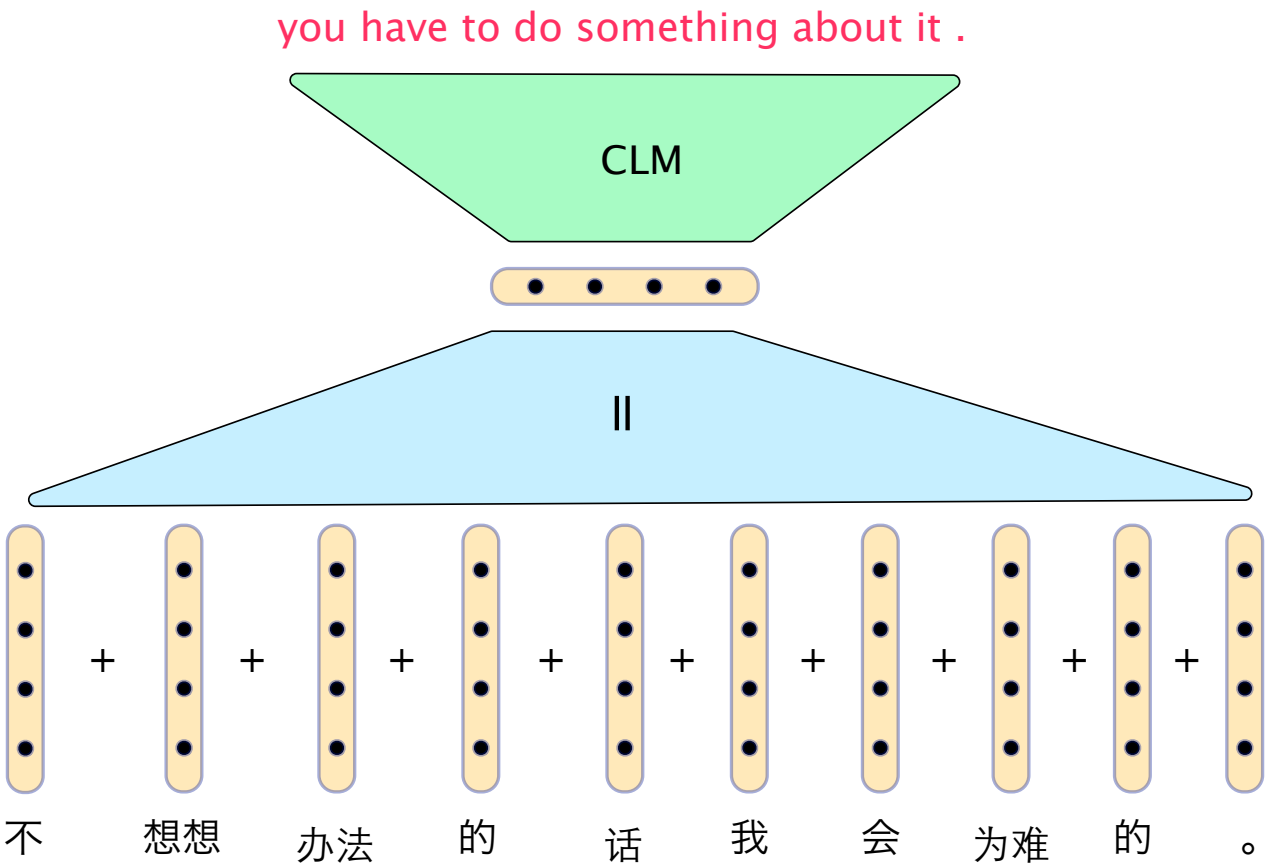
Conditional Generation: A Naive First Model



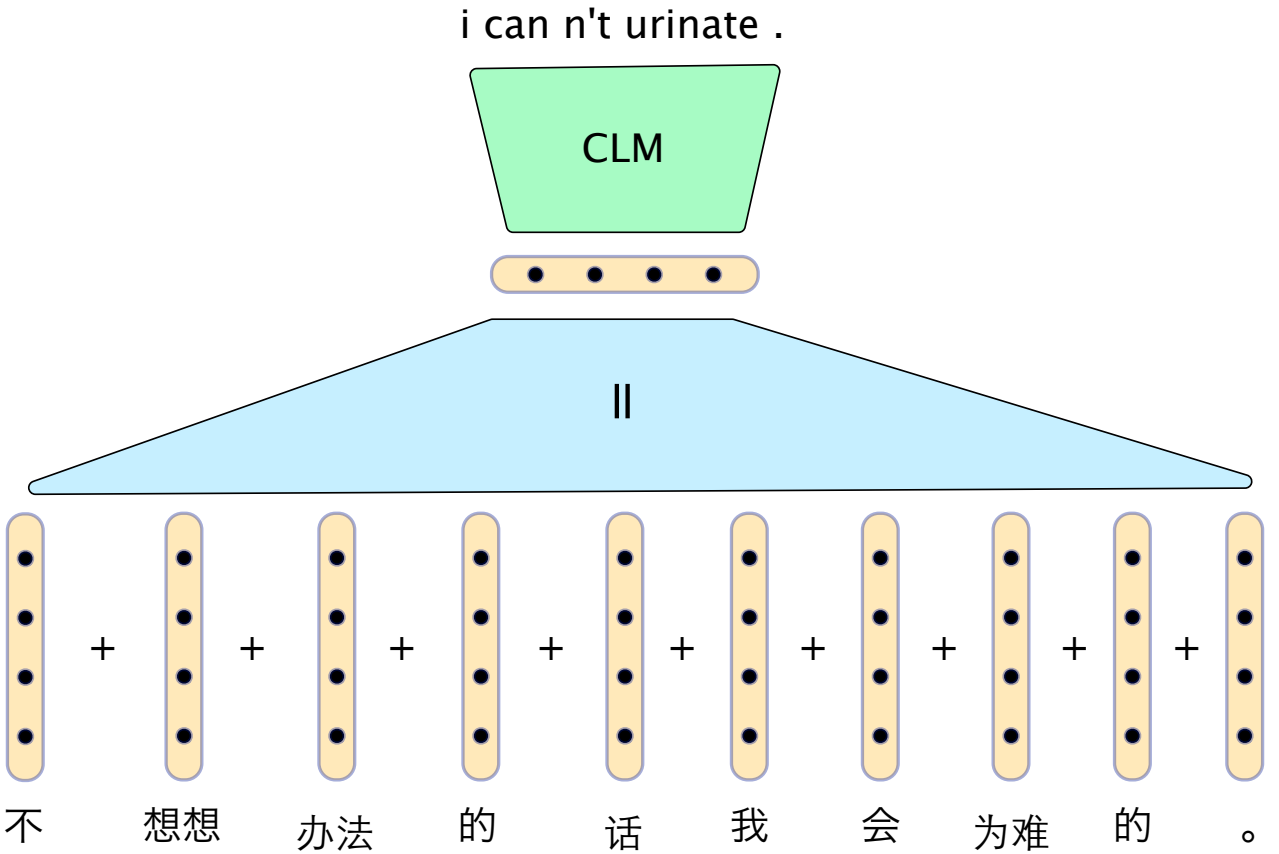
Google Translate

I want a late thirties under $'s room.

Conditional Generation: A Naive First Model



Conditional Generation: A Naive First Model



Conditional Generaton: Small test dataset

Chinese (zh) $\rightarrow$ English (en)	*test 1*	*test 2*
cdec (state-of-the-art MT)	50.1	58.9
Direct (naive bag of words source)	30.8	33.2
Direct (convolution $p(en\|zh)$)	44.6	50.4
Noisy Channel (convolution $p(zh\|en)p(en)$)	50.1	51.8
Noisy Channel $\times$ Direct	51.0	55.2

BLEU score results on a small Chinese $\rightarrow$ English translation task.

Summary

Advantages

- unsupervised features extraction alleviates domain and language dependencies.
- very compact models.
- distributed representations for words naturally include morphological properties.
- the conditional generation framework easily permits additional context such as dialogue and domain level vectors.

Challenges

- better conditioning on sentence position for long sentences, and all the other things this model does not capture!
- handling rare and unknown words.

References

Compositional Morphology for Word Representations and Language Modelling
Botha and Blunsom. ICML 2014

Multilingual Models for Compositional Distributed Semantics
Hermann and Blunsom. ACL 2014

A Convolutional Neural Network for Modelling Sentences
Kalchbrenner, Grefenstette, and Blunsom. ACL 2014

Recurrent Continuous Translation Models
Kalchbrenner and Blunsom. EMNLP 2013

The Role of Syntax in Vector Space Models of Compositional Semantics
Hermann and Blunsom. ACL 2013

Computational Linguistics at The University of Oxford

UNIVERSITY OF OXFORD

DEPARTMENT OF COMPUTER SCIENCE

We are growing!
Postdoctoral and DPhil studentships are available working in Machine Learning and Computational Linguistics

`http://www.clg.ox.ac.uk`