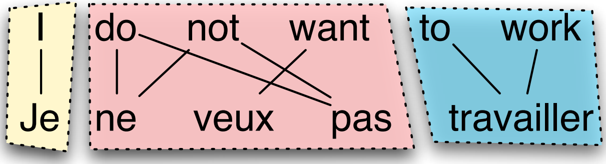
Compositional Semantics, Deep Learning, and Machine Translation

Karl Moritz Hermann, Nal Kalchbrenner, and **Phil Blunsom**

phil.blunsom@cs.ox.ac.uk

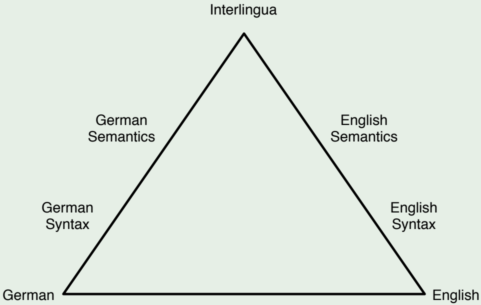
Semantics in MT



- There is not much other than (lexical) semantics in a phrase based MT system.
- What people actually mean when they semantics is often generalisation.

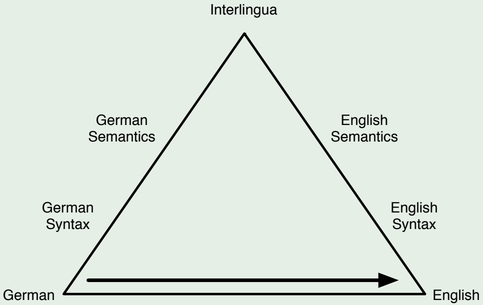
Introduction

The Machine Translation Pyramid:



Introduction

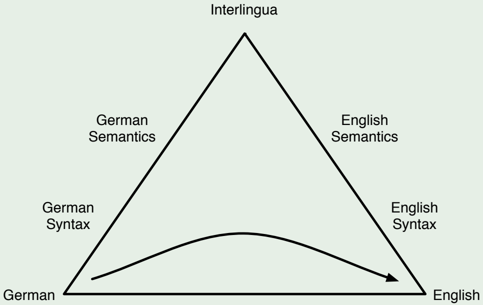
The Machine Translation Pyramid:



Phrase based

Introduction

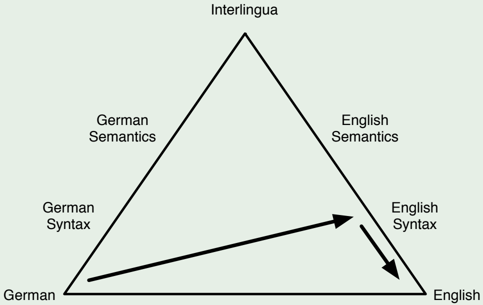
The Machine Translation Pyramid:



Hierarchical (Hiero) MT

Introduction

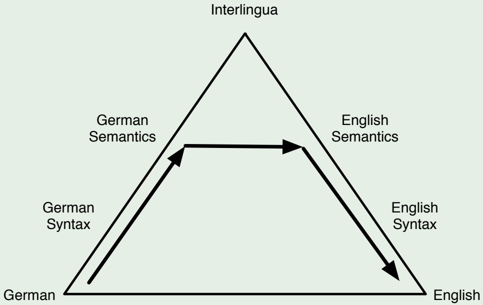
The Machine Translation Pyramid:



String to tree

Introduction

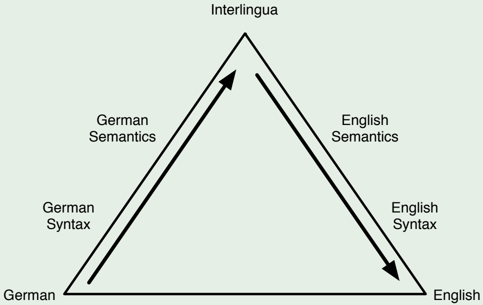
The Machine Translation Pyramid:



Semantic transfer

Introduction

The Machine Translation Pyramid:

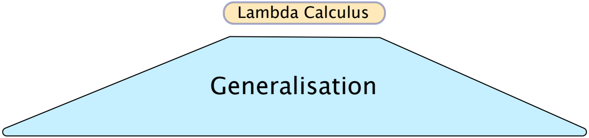


Interlingua: the language of God/lambda calculus

Generalisation in MT

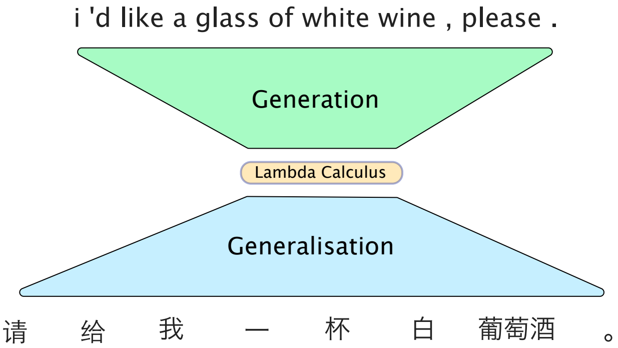
请 给 我 一 杯 白 葡萄酒 。

Generalisation in MT

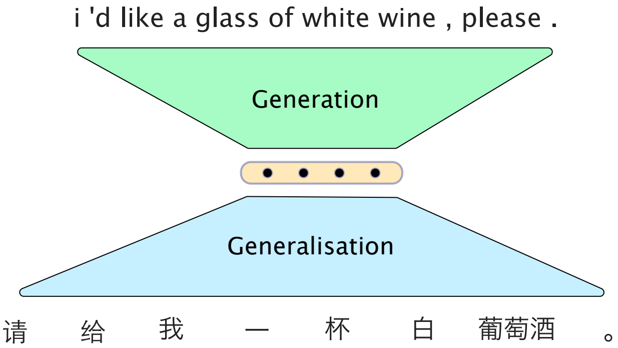


请 给 我 一 杯 白 葡萄酒 。

Generalisation in MT



Generalisation in MT



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

Outline

1. Distributed Representations in Compositional Semantics

2. From Vector Space Compositional Semantics to MT

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 simple task

Q: Do two words (roughly) mean the same?
"Cat" $\equiv$ "Dog" ?

A: Use a distributional representation to find out.

Given a vector representation, we can calculate the similarity between two things using their cosine. We know that[1]

$$A \cdot B = \|A\| \|B\| cos(\theta)$$

Where $cos(\theta)$ is the cosine of the angle between the two vectors A and B. From this it follows that:

$$Sim(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

[1] http://en.wikipedia.org/wiki/Cosine_similarity

Word-Word Similarity

$cos(\theta)$ lies on a range between -1 and 1, with 1 indicating full similarity and 0 indicating no relation and -1 indicating exact opposites.

Cat	Dog
$\begin{bmatrix} 0.7 \\ 0.6 \\ 0.6 \end{bmatrix}$	$\begin{bmatrix} 0.7 \\ -0.3 \\ 0.1 \end{bmatrix}$

$Sim(cat, dog) = 0.437$

Villa	House
$\begin{bmatrix} 0.4 \\ 0.5 \\ 0.3 \end{bmatrix}$	$\begin{bmatrix} 0.3 \\ 0.4 \\ 0.2 \end{bmatrix}$

$Sim(villa, house) = 0.998$

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 a distributional representation to find out?

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 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

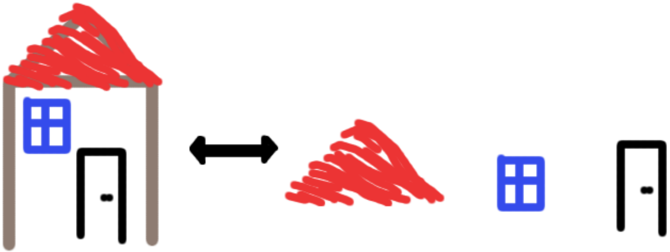
- One word per entry: Size of dictionary (small)
- One sentence per entry: Number of possible sentences (infinite)

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

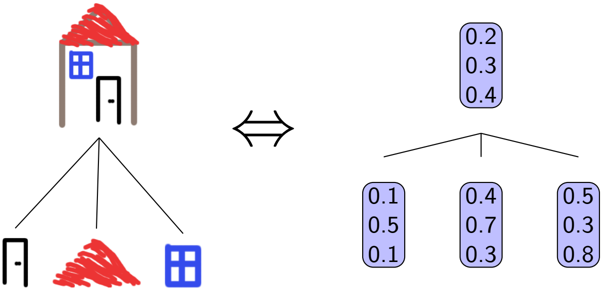
What is Deep Learning

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

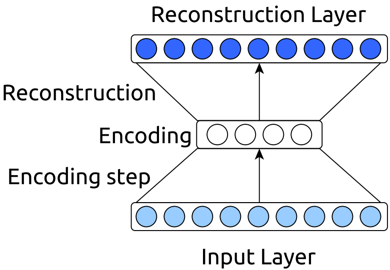
- Addition
- Multiplication
- Kronecker Product
- Tensor Magic
- Matrix-Vector multiplication
- ...

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:

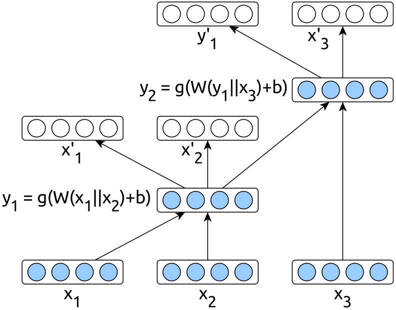


For this to work, our autoencoder minimizes an objective function over inputs $x_i, i \in N$ and their reconstructions x_i':

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

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
$g(W * input + bias)$

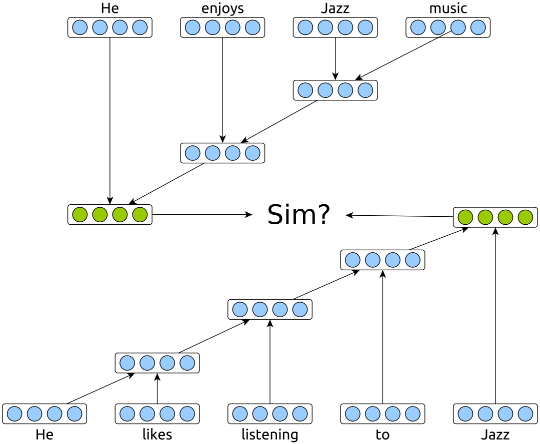
g is a non-linearity (tanh, sigm)
W is a weight matrix
b is a bias

A different task: paraphrase detection

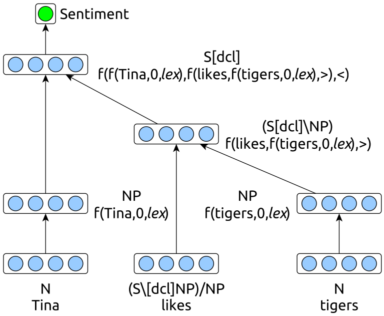
Q: Do two sentences (roughly) mean the same?

"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

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

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

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

2. State of the art for a number of tasks:

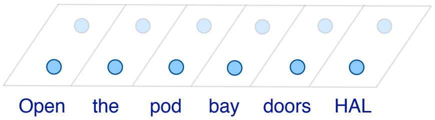
Sentiment Analysis

Paraphrase Detection

Image Search

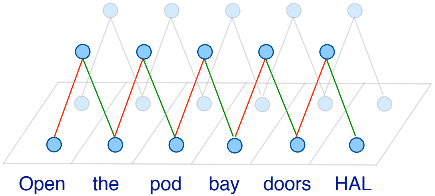
...

A Convolution Sentence Model

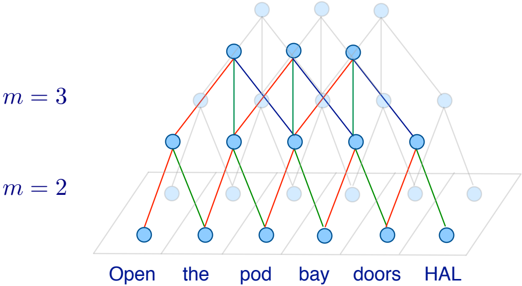


A Convolution Sentence Model

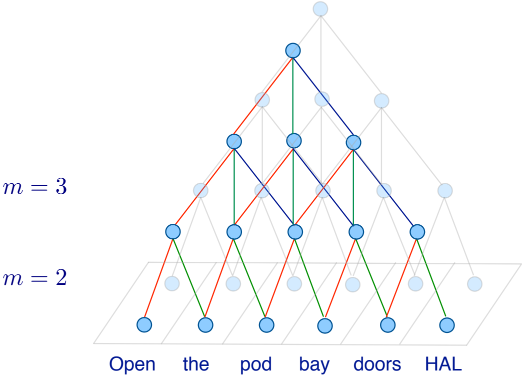
$m = 2$



A Convolution Sentence Model



A Convolution Sentence Model



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.

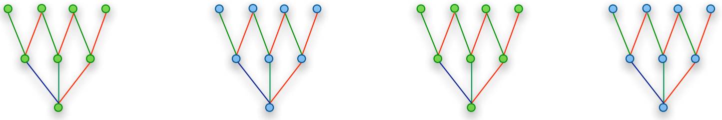
A CSM for Dialogue Act Tagging

Dave: Hello HAL, do you read me HAL?

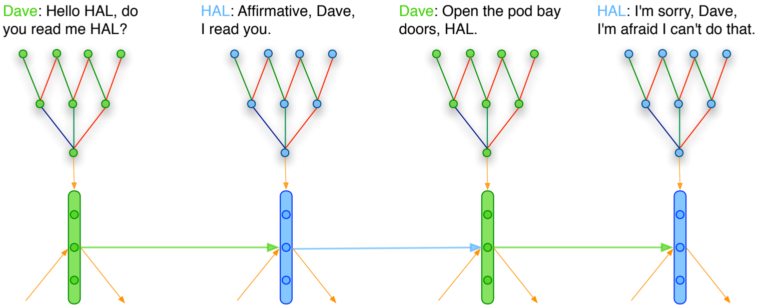
HAL: Affirmative, Dave, I read you.

Dave: Open the pod bay doors, HAL.

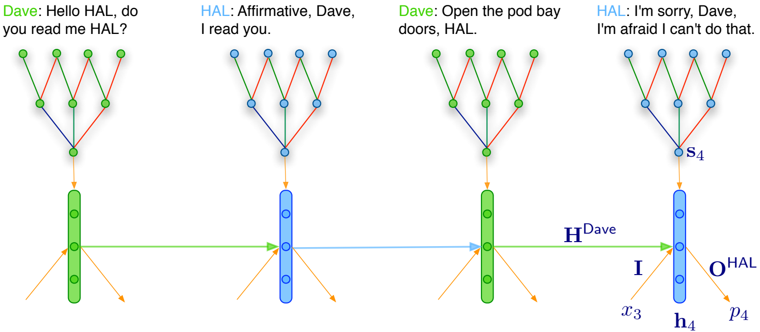
HAL: I'm sorry, Dave, I'm afraid I can't do that.



A CSM for Dialogue Act Tagging



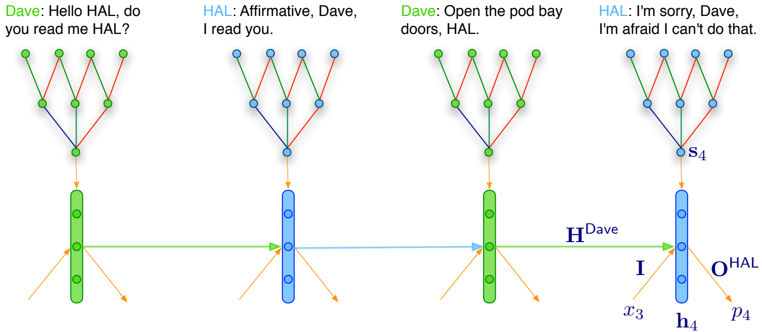
A CSM for Dialogue Act Tagging



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

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

A CSM for Dialogue Act Tagging



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

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

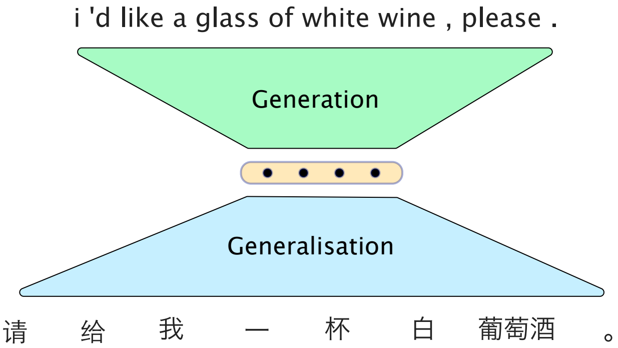
State of the art results while allowing online processing of dialogue.

Outline

1 Distributed Representations in Compositional Semantics

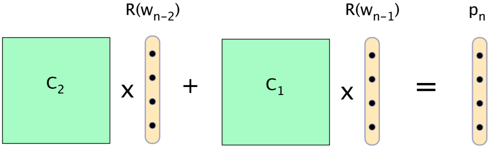
2 From Vector Space Compositional Semantics to MT

Generalisation in MT



Generation

A simple distributed representation language model:

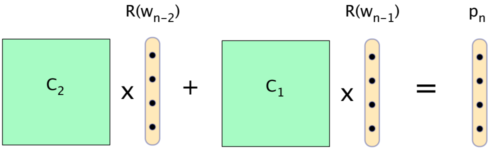


$$p_n = C_{n-2} R(w_{n-2}) + C_{n-1} R(w_{n-1})$$
$$p(w_n | w_{n-1}, w_{n-2}) \propto \exp\left(R(w_n)^T p_n\right)$$

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

Generation

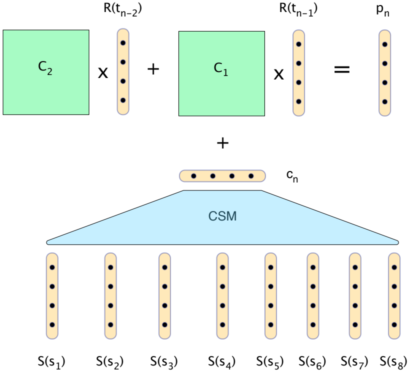
A simple distributed representation language model:



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

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

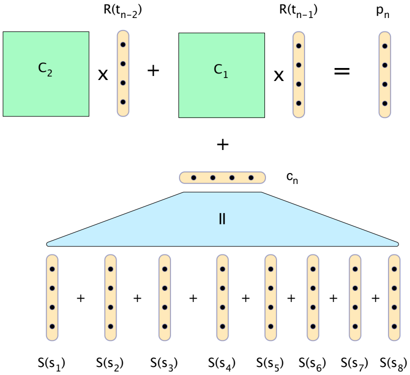
Conditional Generation



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

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

Conditional Generation: A Naive First Model



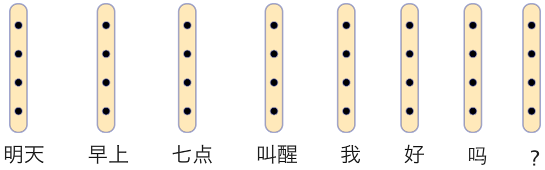
$$p_n = C_2 R(t_{n-2}) + C_1 R(t_{n-1}) + \sum_{j=1}^{|\mathbf{s}|} S(s_j)$$

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

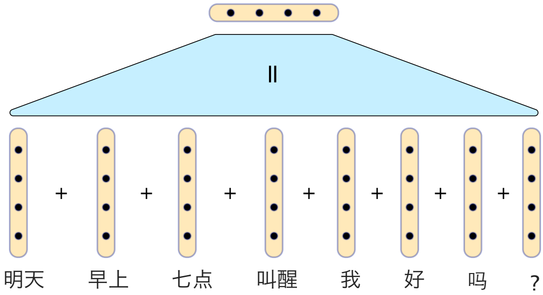
Conditional Generation: A Naive First Model

明天 早上 七点 叫醒 我 好 吗 ?

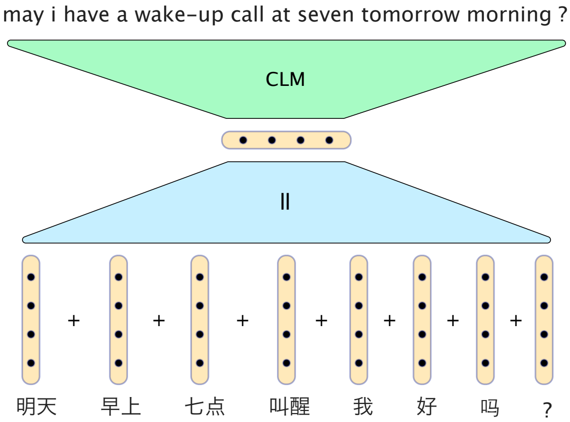
Conditional Generation: A Naive First Model



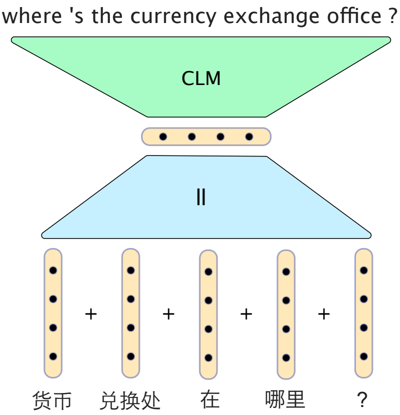
Conditional Generation: A Naive First Model



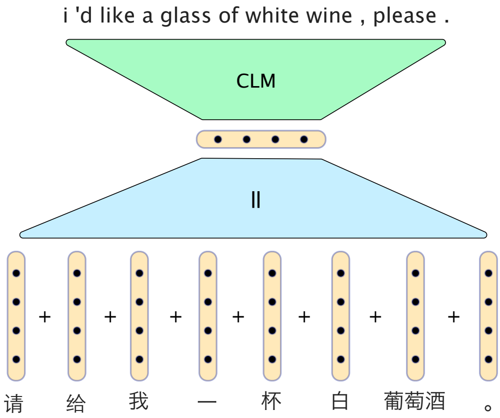
Conditional Generation: A Naive First Model



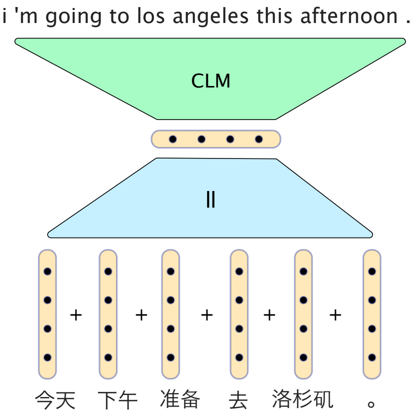
Conditional Generation: A Naive First Model



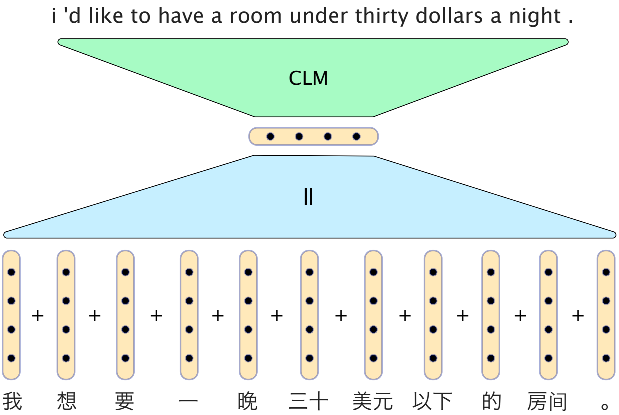
Conditional Generation: A Naive First Model



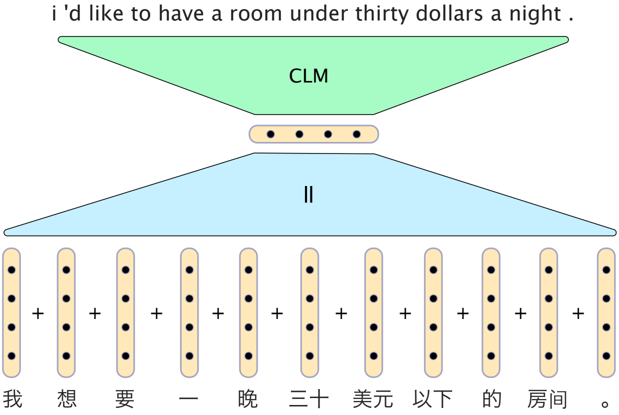
Conditional Generation: A Naive First Model



Conditional Generation: A Naive First Model



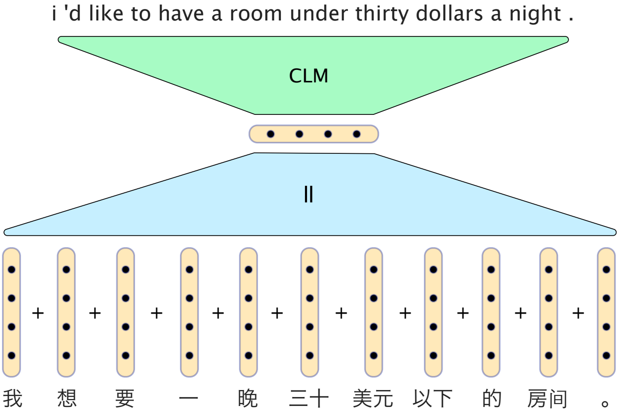
Conditional Generation: A Naive First Model



Rough Gloss

I would like a night thirty dollars under room.

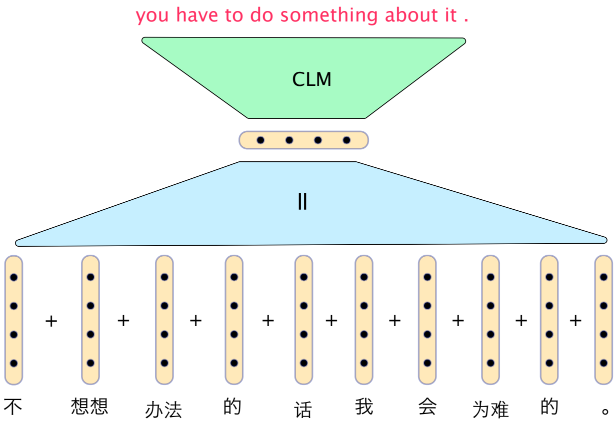
Conditional Generation: A Naive First Model



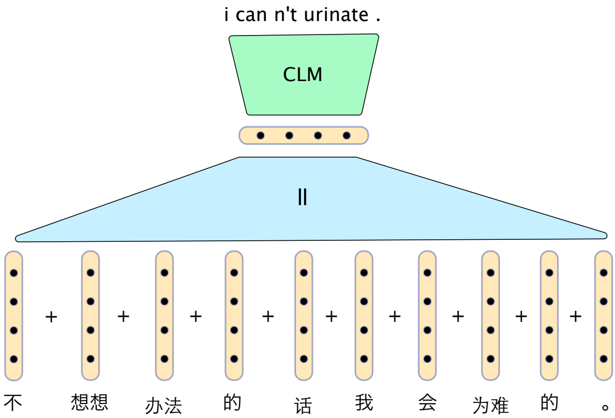
Google Translate

I want a late thirties under $'s room.

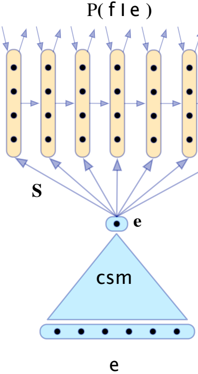
Conditional Generation: A Naive First Model



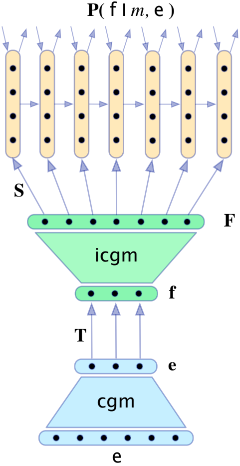
Conditional Generation: A Naive First Model



Conditional Generation: A Convolution N-Gram Model



RCTM I



RCTM II

A Convolution N-Gram Model

En → Fr	*2009*	*2010*	*2011*	*2012*
Knesser-Ney 5gram	218	213	222	225
RNNLM	178	169	178	181
IBM Model 1	207	200	188	197
fast_align (cdec/IBM Model 2)	153	146	135	144
RCTM I	143	134	140	142
RCTM II	**86**	**77**	**76**	**77**

Perplexity results on the WMT News-Commentary test sets.

A Convolution N-Gram Model

En → Fr	*2009*	*2010*	*2011*	*2012*
Knesser-Ney 5gram	218	213	222	225
RNNLM	178	169	178	181
IBM Model 1	207	200	188	197
fast_align (cdec/IBM Model 2)	153	146	135	144
RCTM I	143	134	140	142
RCTM II	**86**	**77**	**76**	**77**

Perplexity results on the WMT News-Commentary test sets.

In k-best rescoring experiments the RCTM II model achieves similar Bleu scores to a MERT trained cdec baseline.

Summary

Advantages

- fast to train and decode, very compact models.
- a valid and tractable probability distribution over translations, making extensions easy to implement.
- 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 and length.
- handling rare and unknown words.

Computational Linguistics at The University of Oxford

UNIVERSITY OF OXFORD

DEPARTMENT OF COMPUTER SCIENCE

Funded studentships are available for strong students interested in pursuing graduate study in Machine Learning and Computational Linguistics

`http://www.cs.ox.ac.uk/admissions/dphil/`