Natural Language Processing
Phrase-based Machine Translation, etc.

Christopher Manning

Borrows slides Kevin Knight and Dan Klein

Feature gains

- The core numeric features should get you a decent baseline MT system
- Expect and be pleased by getting small incremental gains from features you devise
- 0.25 BLEU from a feature is good
- 0.5 BLEU from a feature is fantastic

Phrase-Based Translation Overview

Input:	lo haré | rápidamente | .	
Translations:	I'll do it | quickly | .	
	quickly | I'll do it | .	

The decoder…

tries different segmentations,

translates phrase by phrase,

and considers reorderings.

Phrase-Based Translation

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. ”
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

Phrase-Based Translation

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. ”
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

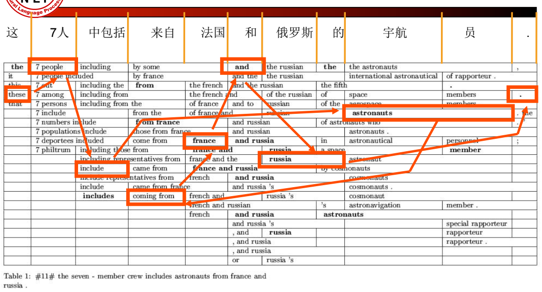
Phrase-Based Translation

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. ”
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

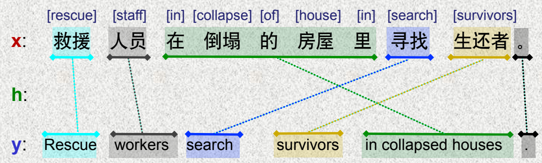
Phrase-Based Translation



Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

Local syntax in phrase-based systems

[Och et al., 1999; Och and Ney; 2004]

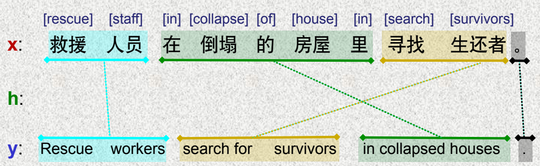


Phrases capture multi-word expressions,
help select correct function words,
and enable local reorderings.

8

Local syntax in phrase-based systems

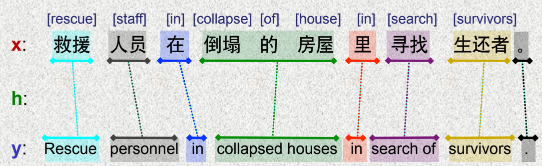
[Och et al., 1999; Och and Ney; 2004]



Phrases capture multi-word expressions,
help select correct function words (e.g., now also "for"),
and enable local reorderings.

9

Phrase-based models at test time



Google translate 's actual output, 2010

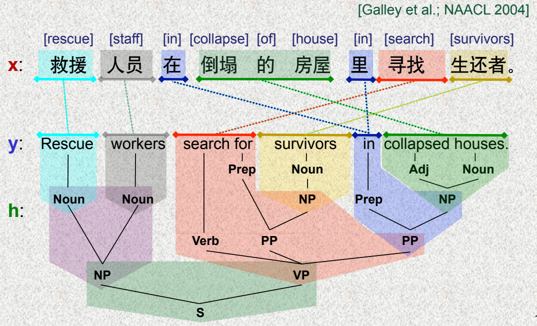
Oct 2013 output: Rescue workers in collapsed buildings in search of survivors.

Long test phrases are often unseen in training.
Short phrases yield poor translations.
Need a more effective model to account for non-local dependencies.

10

Syntax-based MT: Translation as parsing

[Galley et al.; NAACL 2004]



11

Natural Language Processing

Language Models

Christopher Manning
CS224N

Language Models

- Traditional grammars (e.g., regular, context free) give a hard ("categorical") model of the sentences in a language
- For NLP, and other applied work, a probabilistic model of a language is *much* more useful
  - It says what people usually say (next)
  - It enables more fine-grained prediction and inference
- Called a Language Model … strange but standard

Uses of language models

- Speech recognition
  - "I saw a van" is a more likely sentence than "eyes awe of an"
- OCR & Handwriting recognition
  - More probable sentences are more likely correct readings
- Machine translation
  - More likely sentences are probably better translations
- (Fluent Text) Generation
  - More likely sentences are probably better NL generations
- Context sensitive spelling correction
  - "Their are problems wit this sentence."
- Predictive text input systems
  - Please turn your cell phone of_
  - Google query completion suggestions

Uses of language models

- Text classification
  - A topic is a language ("the language of finance")
- Gender/style detection
- Information retrieval: Language models for IR
  - Treat either or both the query or each document as a "language"
  - One of the most pursued research approaches recently (at UMass/CMU)
- Certain aspects of grammar checking
  - E.g., preposition choice
- Text compression
  - Way better than gzip/bzip2 for human language text

Probabilistic Language Models

- Idea is to build models which assign scores to sentences
  - P(I saw a van) >> P(eyes awe of an)
  - Not really grammaticality
    - P(artichokes intimidate zippers) ≈ 0
- Formally, a probability distribution over sentences of a language … sums to 1 over whole language
- One option: empirical distribution over corpus sentences?
  - Problem: doesn't generalize (at all)
  - Whereas languages are infinite

Probabilistic Language Models

Three major components of generalization

- ***Decomposition***: sentences generated in small steps
- ***Discounting***: save some probability mass for the possibility of unseen events
- ***Backoff*** contexts that words are generated from to equivalence classes of contexts which generalize better

After that, there are a lot of details

- But the details are ***very*** important in getting good performance in many NLP systems

Decomposition: N-Gram Language Models

- No loss of generality to break sentence probability down with the chain rule

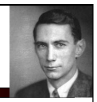
$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

- Too many histories!
  - P(??? | No loss of generality to break sentence) ?
  - P(??? | the water is so transparent that) ?
- N-gram solution: assume each word depends only on a short linear history (a Markov assumption) = equivalence classing

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1})$$

$$= \prod_i P(w_i \mid w_{i-1}) \text{ for bigram}$$

Character-level



- Claude Shannon (1951): the entropy of English
  - http://www.math.ucsd.edu/~crypto/java/ENTROPY/

$$H(X) = E_P \log \frac{1}{P(X)}$$
$$= -\sum_{x \in \mathcal{X}} P(x) \log P(x)$$

- Cross entropy

$$H(S \mid M) = \frac{-\log_2 P_M(S)}{|S|} = \frac{-\sum_{i=1,\ldots,N} \log_2 P_M(w_i \mid w_{1,\ldots,i-1})}{N}$$

e.g.,

$$\sum_j \log_2 P_M(w_j \mid w_{j-1})$$

Word level

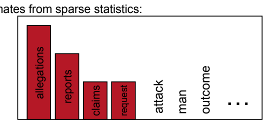
… denied the ________

The Palestinian security chief in Gaza **denied the** report
Judge Kathleen Kennedy-Powell **denied the** motion to strike
Pineau-Valencienne has **denied the** charges
The FDA **denied the** group's request
the show's writer and co-star, **denied the** characters had real-life
The district attorney's office had **denied the** KCBS-TV report
Coleman **denied the** charge
Defense attorney Al Kitching **denied the** allegations
Local officials have consistently **denied the** existence of armed
Kraft has categorically **denied the** remarks
Goddard has **denied the** charges
congressional employees are **denied the** legal protections
who **denied the** accusation of the woman

Discounting/Smoothing

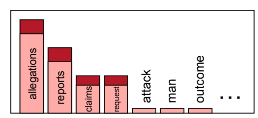
- We often want to make estimates from sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total



- Smoothing flattens spiky distributions so they generalize better

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



- Very important all over NLP, but easy to do badly!
- Illustration with trigrams (h = previous word, could be anything).

Discounting/Backoff/Interpolation

- P(w_i|h) is just a multinomial
  - but we need to estimate it well
  - We want to know how often a word follow some history *h*
  - There's some true distribution P(*w* | *h*)
  - We saw some small sample of *N* words from P(*w* | *h*)
  - We want to reconstruct a useful approximation of P(*w* | *h*)
  - Counts of events we didn't see are always too low
  - Counts of events we did see are *in aggregate* too high
- Discounting: providing mass for what we haven't seen
- Backoff: Increasing *N* by decreasing the amount of history *h*
- Interpolation between backed-off distributions: how to allocate that mass amongst unseen events

Evaluation

- What we want to know is:
  - Will our language model prefer natural sentences?
    - Does it assign *higher probability* to "real" or "frequently observed" sentences than "rarely observed" sentences?
- We train parameters of our model on a **training set**.
- To evaluate how well our model works, we look at the model's performance on some **different** data
- This is what happens in the real world; we want to know how our model performs on data we haven't seen
- So a **test set**. A dataset which is different from our training set
  - Preferably totally unseen/unused!
- So we can do this, we do model development with a separate development test (**devtest**) set

Language models

- Language models are a cool technology
- You can have them for not only a language like "English" but for particular languages/topics
  - Papers about language modeling
  - "Spam emails"
  - Seventeenth century novels
- Because they flexibly model higher order context, they can be very powerful models
  - And work very well

Look at the videos and J&M chapter 4!