Natural Language Processing
Phrase-based Machine Translation, etc.

Christopher Manning

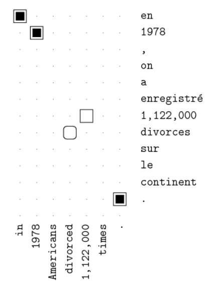
Borrows slides Kevin Knight and Dan Klein

Evaluating Alignments: Alignment Error Rate (Och & Ney 2000)

- ☐ = Sure
- ○ = Possible
- ■ = Alignments (predicted)

$$AER(A,S,P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$

$$= \left(1 - \frac{3+3}{3+4}\right) = \frac{1}{7}$$



Most work has used AER and we do, but it is problematic, and it's better to use an alignment F measure (Fraser and Marcu 2007)

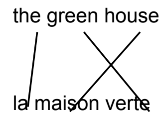
Comparative results (AER)

[Och & Ney 2003]

Model	Training scheme	Size of training corpus 0.5K	8K	128K	1.47M
Dice		50.9	43.4	39.6	38.9
Dice+C		46.3	37.6	35.0	34.0
Model 1	1^5	40.6	33.6	28.6	25.9
Model 2	$1^5 2^5$	46.7	29.3	22.0	19.5
HMM	$1^5 H^5$	26.3	23.3	15.0	10.8
Model 3	$1^5 2^5 3^3$	43.6	27.5	20.5	18.0
	$1^5 H^5 3^3$	27.5	22.5	16.6	13.2
Model 4	$1^5 2^5 3^3 4^3$	41.7	25.1	17.3	14.1
	$1^5 H^5 3^3 4^3$	26.1	20.2	13.1	9.4
	$1^5 H^5 4^3$	26.3	21.8	13.3	9.3
Model 5	$1^5 H^5 4^3 5^3$	26.5	21.5	13.7	9.6
	$1^5 H^5 3^3 4^3 5^3$	26.5	20.4	13.4	9.4
Model 6	$1^5 H^5 4^3 6^3$	26.0	21.6	12.8	8.8
	$1^5 H^5 3^3 4^3 6^3$	25.9	20.3	12.5	8.7

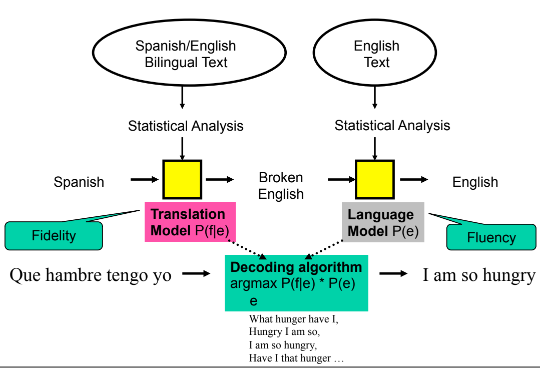
Common software: GIZA++/Berkeley Aligner

Alignments: linguistics



- There isn't enough linguistics to explain this in the translation model … have to depend on the language model … that may be unrealistic … and may be harming our translation model

A Division of Labor



Getting Parallel Sentence Data

- Expensive way:
  - Pay people to translate stuff
- Pretty hard way: Find it, and then earn it
  - Crawl web identifying likely parallel text (or use CommonCrawl)
  - Do a lot of work with formatting, character encodings, doc regions
- Easy way: Use existing data
  - Linguistic Data Consortium (LDC)
    - http://www.ldc.upenn.edu/
    - ~200 million words for some pairs (e.g., Chinese-English)
  - EuroParl:
    - http://www.statmt.org/europarl/
    - Around 50 million words per language for "old" EU countries

Sentence Alignment

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Done by similar Dynamic Programming or EM: see FSNLP ch. 13 for details

Search for Best Translation

voulez – vous vous taire !

Search for Best Translation

voulez – vous vous taire !

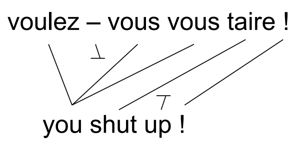
you – you you quiet !

Search for Best Translation

voulez – vous vous taire !

quiet you – you you !

Search for Best Translation



Searching for a translation

Of all conceivable English word strings, we want the one maximizing P(e) x P(f | e)

Exact search

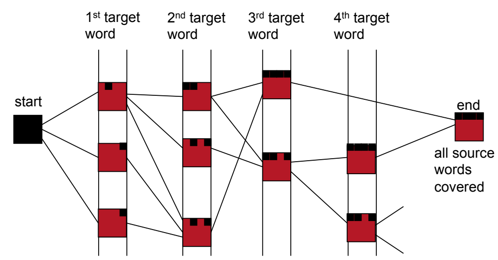
- Even if we have the right words for a translation, there are **n!** permutations.
- We want the translation that gets the highest score under our model
- Finding the `argmax` with a n-gram language model is NP-complete [Germann et al. 2001].
- Equivalent to Traveling Salesman Problem

14

Searching for a translation

- Several search strategies are available
  - Usually a beam search where we keep multiple stacks for candidates covering the same number of source words
  - Or, we could try "greedy decoding", where we start by giving each word its most likely translation and then attempt a "repair" strategy of improving the translation by applying search operators (Germann et al. 2001)
- Each potential English output is called a *hypothesis*.

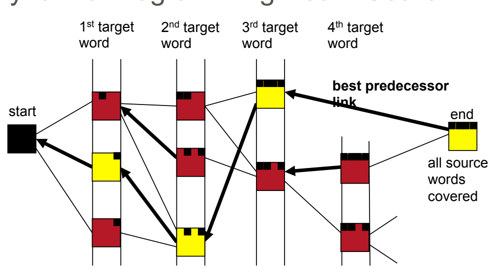
Dynamic Programming Beam Search



Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969; Brown et al, 1996 US Patent; (Och, Ueffing, and Ney, 2001]

Dynamic Programming Beam Search



Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969; Brown et al, 1996 US Patent; (Och, Ueffing, and Ney, 2001]

MT Evaluation

Illustrative translation results

- *la politique de la haine .* *(Foreign Original)*
- politics of hate . (Reference Translation)
- the policy of the hatred . (IBM4+N-grams+Stack)

- *nous avons signé le protocole .* *(Foreign Original)*
- we did sign the memorandum of agreement . (Reference Translation)
- we have signed the protocol . (IBM4+N-grams+Stack)

- *où était le plan solide ?* *(Foreign Original)*
- but where was the solid plan ? (Reference Translation)
- where was the economic base ? (IBM4+N-grams+Stack)

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

the Ministry of Foreign Trade and Economic Cooperation, including foreign
direct investment 40.007 billion US dollars today provide data include
that year to November china actually using foreign 46.959 billion US dollars and

MT Evaluation

- Manual (the best!?):
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - **Adequacy and Fluency** (5 or 7 point scales)
  - Error categorization
  - **Comparative ranking of translations**
- Testing in an application that uses MT as one sub-component
  - E.g., question answering from foreign language documents
    - May not test many aspects of the translation (e.g., cross-lingual IR)
- Automatic metric:
  - WER (word error rate) – why problematic?
  - **BLEU (Bilingual Evaluation Understudy)**

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to match same portion of reference translation twice at a certain n-gram level (two MT words *airport* are only correct if two reference words *airport;* can't cheat by typing out "the the the the the")
  - Do count unigrams also in a bigram for unigram precision, etc.
- Brevity Penalty
  - Can't just type out single word "the" (precision 1.0!)
- It was thought quite hard to "game" the system (i.e., to find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it's precision-oriented
- BLEU4 formula
  (counts n-grams up to length 4)

exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Note: only works at corpus level (zeroes kill it); there's a smoothed variant for sentence-level

BLEU in Action

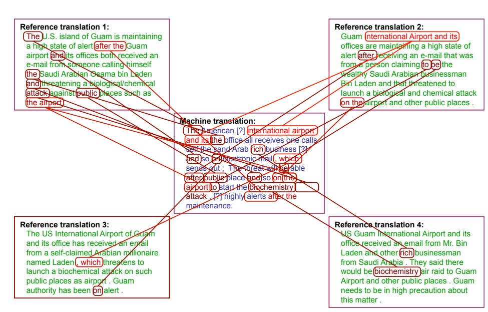
枪手被警方击毙。 **(Foreign Original)**

the gunman was shot to death by the police . **(Reference Translation)**

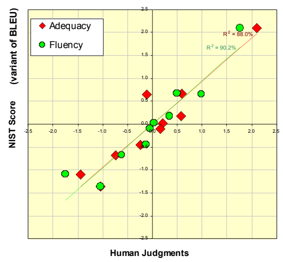
the gunman was police kill . **#1**
wounded police jaya of **#2**
the gunman was shot dead by the police . **#3**
the gunman arrested by police kill . **#4**
the gunmen were killed . **#5**
the gunman was shot to death by the police . **#6**
gunmen were killed by police ?SUB>0 ?SUB>0 **#7**
al by the police . **#8**
the ringer is killed by the police . **#9**
police killed the gunman . **#10**

green = 4-gram match (good!)
red = word not matched (bad!)

Multiple Reference Translations



Initial results showed that BLEU predicts human judgments well



slide from G. Doddington (NIST)

Automatic evaluation of MT

- People started optimizing their systems to maximize BLEU score
  - BLEU scores improved rapidly
  - The correlation between BLEU and human judgments of quality went way, way down
  - StatMT BLEU scores now approach those of human translations but their true quality remains far below human translations
- Coming up with automatic MT evaluations has become its own research field
  - There are many proposals: TER, METEOR, MaxSim, SEPIA, our own RTE-MT
  - TERpA is a representative good one that handles some word choice variation.
- MT research **requires** *some* automatic metric to allow a rapid development and evaluation cycle.

Phrase-Based Statistical MT

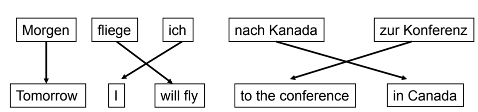
MT Problems to address

- Funny asymmetry of IBM models
- More features for translation quality
- Work with larger chunks than just words
  - Phrase-based systems
- Hey, what about some linguistic structure?
  - Hierarchical and grammar-based systems

Flaws of Word-Based MT

- Multiple English words for one French word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - “real estate”, “note that”, “interested in”
  - There’s a lot of multiword idiomatic language use
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place

Phrase-Based Statistical MT



- Foreign input segmented into phrases
  - “phrase” is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered

See J&M or Lopez 2008 for an intro.

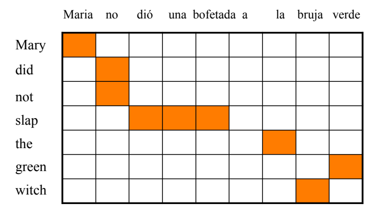
This is still pretty much the state-of-the-art!

Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Local context is very useful for disambiguating
  - "interest rate" → …
  - "interest in" → …
- The more data, the longer the learned phrases
  - Sometimes whole sentences

How to Learn the Phrase Translation Table?

- Main method: "alignment templates" (Och et al, 1999)
- Start with "symmetrized" word alignment, build phrases from that.

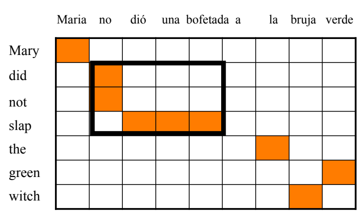


This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.

This is the best (or "Viterbi") alignment.

How to Learn the Phrase Translation Table?

- One method: "alignment templates" (Och et al, 1999)
- Start with word alignment, build phrases from that.

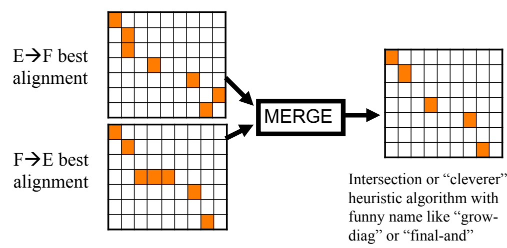


This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.

This is the best (or "Viterbi") alignment.

IBM Models are 1-to-Many

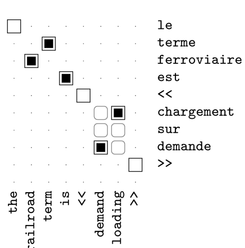
- Run IBM-style aligner both directions, then merge:



Intersection or "cleverer" heuristic algorithm with funny name like "grow-diag" or "final-and"

Symmetrization

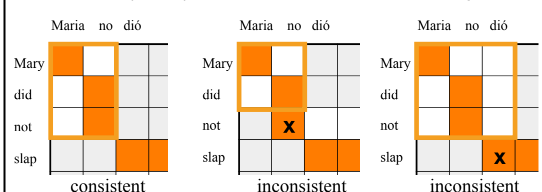
- Standard practice is to train models in each direction then to intersect their predictions
- Second model is basically a filter on the first
  - Precision jumps, recall drops
  - End up not guessing hard alignments



Model	P/R	AER
Model 1 E→F	82/58	30.6
Model 1 F→E	85/58	28.7
Model 1 AND	96/46	34.8

How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*



- Phrase alignment must contain all alignment points for all the words in both phrases!
- These phrase alignments are sometimes called *beads*

The phrase table becomes our translation model. How do we put goodness values on phrases?

```
开发 ||| the development ||| (1) ||| () (0) ||| -3.43 -2.72 -3.43 -2.76
开发 ||| the development of ||| (1) ||| () (0) () ||| -4.03 -2.72 -4.26 -5.31
开发 ||| development ||| (0) ||| (0) ||| -2.97 -2.72 -0.86 -0.95
开发 ||| development of ||| (0) ||| (0) () ||| -3.41 -2.72 -3.22 -3.50
进行 监督 ||| that carries out a supervisory ||| (1,2,3) (4) ||| () (0) (0) (0) (1) ||| 0.0 -3.68 -7.27 -21.24
进行 监督 ||| carries out a supervisory ||| (0,1,2) (3) ||| (0) (0) (0) (1) ||| 0.0 -3.68 -7.27 -17.17
监督 ||| supervisory ||| (0) ||| (0) ||| -1.03 -0.80 -3.68 -3.24
监督 检查 ||| supervisory inspection ||| (0) (1) ||| (0) (1) ||| 0.0 -2.33 -6.07 -4.85
检查 ||| inspection ||| (0) ||| (0) ||| -1.54 -1.53 -2.05 -1.60
尽管 ||| in spite ||| (1) ||| () (0) ||| -0.90 -0.50 -3.56 -6.14
尽管 ||| in spite of ||| (1) ||| () (0) () ||| -1.11 -0.50 -3.93 -8.68
尽管 ||| in spite of the ||| (1) ||| () (0) () () ||| -1.06 -0.50 -4.77 -10.50
尽管 ||| in spite of the fact ||| (1) ||| () (0) () () () ||| -1.18 -0.50 -6.54 -18.19
尽管 ||| spite ||| (0) ||| (0) ||| -0.78 -0.50 -3.34 -2.88
尽管 ||| spite of ||| (0) ||| (0) () ||| -0.96 -0.50 -3.71 -5.43
尽管 ||| spite of the ||| (0) ||| (0) () () ||| -0.90 -0.50 -4.54 -7.25
尽管 ||| spite of the fact ||| (0) ||| (0) () () () ||| -0.99 -0.50 -6.25 -14.93
尽管 ||| spite of the fact that ||| (0) ||| (0) () () () () ||| -1.03 -0.50 -6.35 -19.00
```

The "Fundamental Equation of Machine Translation" (Brown et al. 1993)

$$\hat{e} = \underset{e}{\operatorname{argmax}}\; P(e \mid f)$$

$$= \underset{e}{\operatorname{argmax}}\; P(e) \times P(f \mid e) / P(f)$$

$$= \underset{e}{\operatorname{argmax}}\; P(e) \times P(f \mid e)$$

What StatMT people do in the privacy of their own homes

$$\underset{e}{\operatorname{argmax}}\; P(e \mid f) =$$

$$\underset{e}{\operatorname{argmax}}\; P(e) \times P(f \mid e) / P(f) =$$

$$\underset{e}{\operatorname{argmax}}\; P(e)^{1.9} \times P(f \mid e)$$ … works better!

Which model are you now paying more attention to?

What StatMT people do in the privacy of their own homes

$$\underset{e}{\operatorname{argmax}}\; P(e \mid f) =$$

$$\underset{e}{\operatorname{argmax}}\; P(e) \times P(f \mid e) / P(f)$$

$$\underset{e}{\operatorname{argmax}}\; P(e)^{1.9} \times P(f \mid e) \times 1.1^{length(e)}$$

Rewards longer hypotheses, since these are 'unfairly' punished by P(e)

What StatMT people do in the privacy of their own homes

$$\underset{e}{\operatorname{argmax}}\; \underbrace{P(e)^{1.9} \times P(f \mid e) \times 1.1^{length(e)} \times KS^{3.7} \ldots}$$

Lots of knowledge sources vote on any given hypothesis. Each has a weight

"Knowledge source" = "feature function" = "score component".

Log-linear feature-based MT

argmax_e 1.9×log P(e) + 1.0×log P(f | e) +
1.1× log length(e) + 3.7×KS + …

= $\operatorname{argmax}_e \Sigma_i\, w_i f_i$

So, we have two things:

- "Features" *f*, such as log language model score
- A weight *w* for each feature that indicates how good a job it does at indicating good translations

Numeric Features for Phrases: Log Phrase Pair Probabilities

- A certain phrase pair (f-f-f, e-e-e) may appear many times across the bilingual corpus.
- No EM training
- Simplest features are just relative frequency!
- $$P(\text{f-f-f} \mid \text{e-e-e}) = \frac{\text{count(f-f-f, e-e-e)}}{\text{count(e-e-e)}}$$
- P(e-e-e | f-f-f)
- Model 1 score P(f|e)
- Model 1 score P(e|f)

Other Numeric Features

- log language model score
- amount of “distortion” [reordering] in the translation hypothesis
- Other good ideas….

Categorical Features

- Categorical features are often represented by a symbol (a String)
- Mathematically, they’re a feature whose value is 0 or 1
  - Source phrase contains verb but target phrase doesn’t: TRANS_NO_VERB
  - Source phrase contains period but target phrase doesn’t: TRANS_NO_PERIOD
  - Target phrase contains the word “the”: THE

Feature weights

- **How to set the weights for features?**
  - Done for you, by optimization procedure
  - One way (which we look at later doing NER): maxent (softmax/logistic) models
  - The standard way is “MERT” (minimum error rate training)
  - A more recent proposal is “PRO” (pairwise ranking maxent optimization)
- **But basically you want a small number if feature slightly/doesn’t indicate a good translation on average, big weight if it does**
  - Positive or negative as positive/negative correlated

Feature gains

- The core numeric features should get you a decent 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

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.

Non-Monotonic Phrasal MT