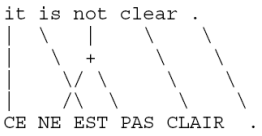
Statistical NLP
Spring 2010

Lecture 18: Phrase / Syntactic MT

Dan Klein – UC Berkeley

Decoding

- First, consider word-to-word models
  - Finding best alignments is easy
  - Finding translations is hard (why?)



Bag "Generation" (Decoding)

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

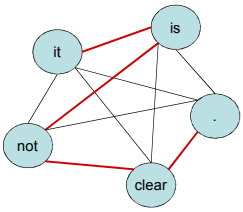
Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

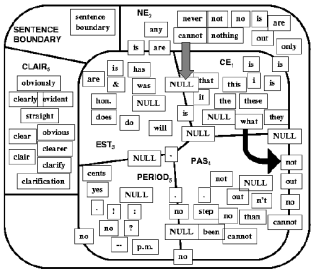
In our organization research has two missions.
⇒ In our missions research organization has two.

Bag Generation as a TSP

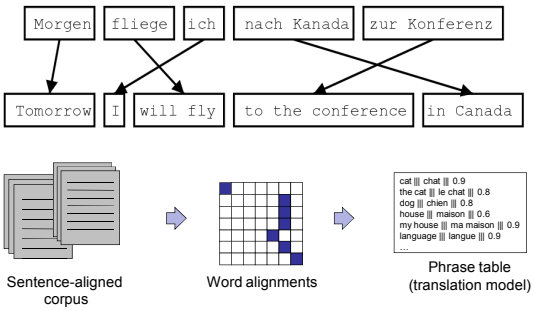
- Imagine bag generation with a bigram LM
  - Words are nodes
  - Edge weights are P(w|w')
  - Valid sentences are Hamiltonian paths
- Not the best news for word-based MT (and it doesn't get better with phrases)



IBM Decoding as a TSP



Phrase-Based Systems



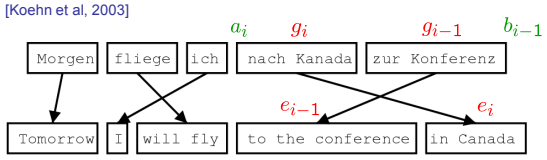
Sentence-aligned corpus

Word alignments

Phrase table (translation model)

The Pharaoh "Model"

[Koehn et al, 2003]



$$P(e|g) = P(\{\bar{g}_i\}|g) \prod_i \phi(\bar{e}_i|\bar{g}_i) d(a_i - b_{i-1})$$

Segmentation Translation Distortion

The Pharaoh "Model"

$$P(f|e) = P(\{\bar{e}_i\}|e) \prod_i \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

$\frac{1}{K}$ $\frac{count(\bar{f}_i, \bar{e}_i)}{count(\bar{e}_i)}$ $\alpha^{|a_i - b_{i-1}|}$

Where do we get these counts?

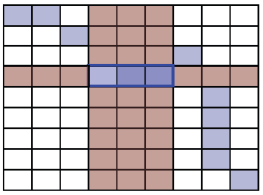
Counting Phrase Pairs

Input:

Gracias , lo haré de muy buen grado .
Thank you , I shall do so gladly .

First, we learn word alignments,

then we infer aligned phrases.



	Gloss
Gracias	*Thanks*
,	*,*
lo	*that*
haré	*do [first; future]*
de	*of*
muy	*very*
buen	*good*
grado	*degree*
.	*.*

Thank you , I shall do so gladly .

Phrase-Based Decoding

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

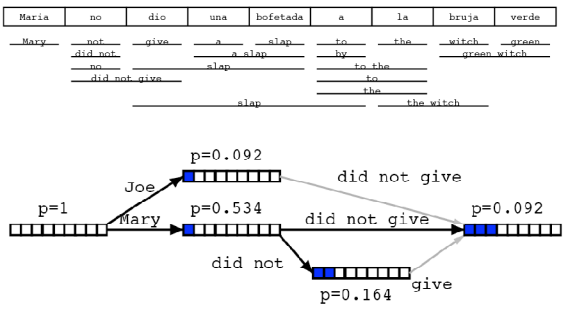
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		.
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**		the french	and the russian		the fifth	.	
these	7 among	including from			the french and	of 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		

Decoder design is important: [Koehn et al. 03]

The Pharaoh Decoder

- Probabilities at each step include LM and TM

Hypotheis Lattices



Pruning

Maria no dio una bofetada a la bruja verde

```
e: Mary did not
f: **--------
p: 0.154
```

better partial translation

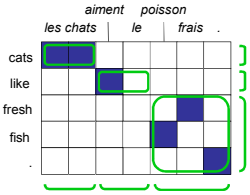
```
e: the
f: -----**--
p: 0.354
```

covers easier part --> lower cost

- Problem: easy partial analyses are cheaper
  - Solution 1: use beams per foreign subset
  - Solution 2: estimate forward costs (A*-like)

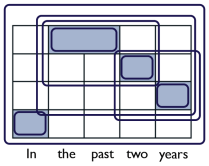
Phrase Scoring

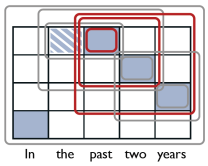
$$\phi_{mle}(\bar{e}_j|\bar{f}_i) = \frac{c(\bar{f}_i, \bar{e}_j)}{c(\bar{f}_i)}$$



- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - … and others
- Seems not to work well, for a variety of partially understood reasons
- Main issue: big chunks get all the weight, obvious priors don't help
  - Though, [DeNero et al 08]

Extraction Sets





过去 [past]
两 [two]
年 [year]
中 [in]

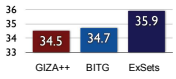
$$\phi(\mathcal{A}) = \sum_{(i,j)\in\mathcal{A}_s} \phi(i,j) + \sum_{[g,h]\Leftrightarrow[k,\ell]\in R_n(\mathcal{A})} \phi(g,h,k,\ell)$$

P(in|中) = 0.8
InDictionary = 1.0

Count(*the past two years*, 过去 两 年) = 7
Size(4,3) = 1

Translation quality for Chinese-to-English

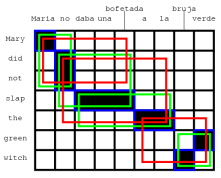
BLEU

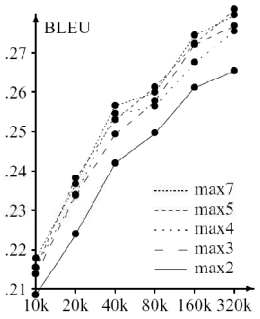


[DeNero and Klein, in submission]

Phrase Size

- Phrases do help
  - But they don't need to be long
  - Why should this be?



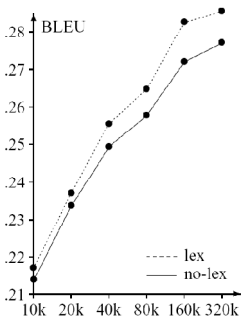


Lexical Weighting

$$\phi(\bar{f}_i|\bar{e}_i) = \frac{count(\bar{f}_i, \bar{e}_i)}{count(\bar{e}_i)} p_w(\bar{f}_i|\bar{e}_i)$$

```
      f1 f2 f3
NULL  -- -- ##
e1    ## -- --
e2    -- ## --
e3    -- ## --
```

$$\begin{aligned} p_w(\bar{f}|\bar{e},a) &= p_w(f_1 f_2 f_3|e_1 e_2 e_3, a) \\ &= w(f_1|e_1) \\ &\times \frac{1}{2}(w(f_2|e_2) + w(f_2|e_3)) \\ &\times w(f_3|\text{NULL}) \end{aligned}$$



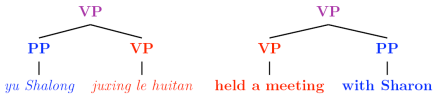
WSD?

- Remember when we discussed WSD?
  - Word-based MT systems rarely have a WSD step
  - Why not?

Syntax-Based MT

- synchronous context-free grammars (SCFGs)
  - context-free grammar in two dimensions
  - generating pairs of strings/trees simultaneously
  - co-indexed nonterminal further rewritten as a unit

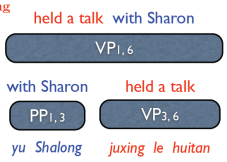
VP → PP(1) VP(2), VP(2) PP(1)
VP → *juxing le huitan*, held a meeting
PP → *yu Shalong*, with Sharon



Translation by Parsing

- translation with SCFGs => monolingual parsing
- parse the source input with the source projection
  - build the corresponding target sub-strings in parallel

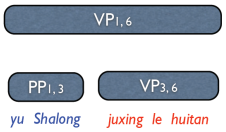
VP → PP(1) VP(2), VP(2) PP(1)
VP → *juxing le huitan*, held a meeting
PP → *yu Shalong*, with Sharon



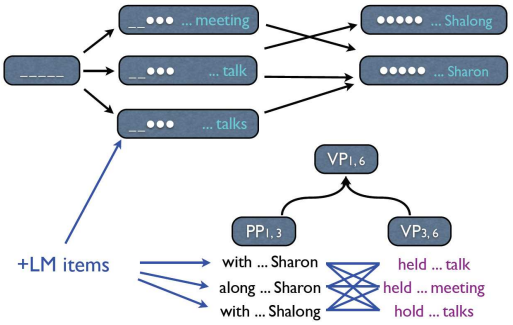
Translation by Parsing

- translation with SCFGs => monolingual parsing
- parse the source input with the source projection
  - build the corresponding target sub-strings in parallel

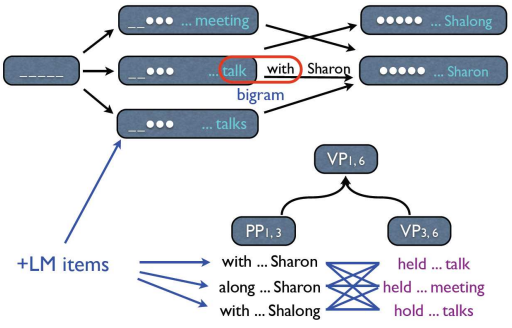
VP → PP(1) VP(2), VP(2) PP(1)
VP → *juxing le huitan*, held a meeting
PP → *yu Shalong*, with Sharon



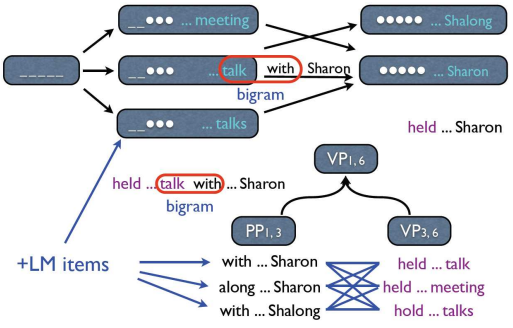
Compact Forests



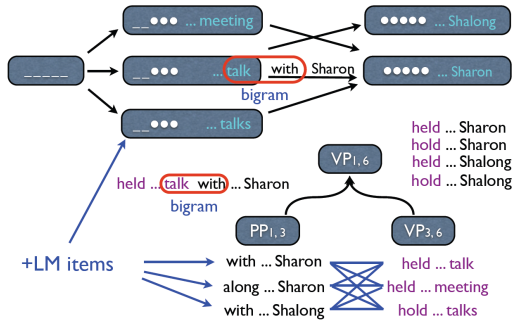
Compact Forests



Compact Forests

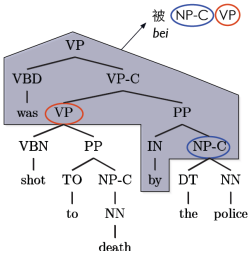


Compact Forests



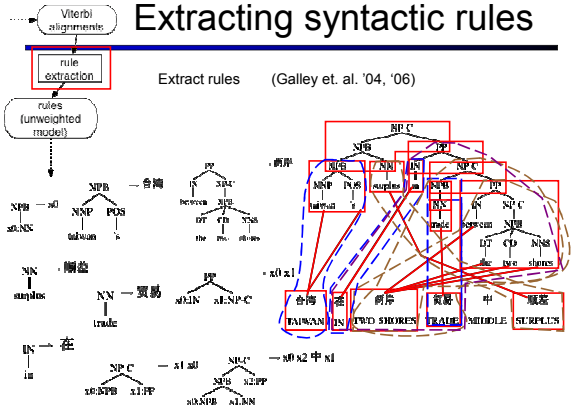
Learning MT Grammars

- syntax-directed, English to Chinese (Huang, Knight, Joshi, 2006)
- first parse input, and then recursively transfer

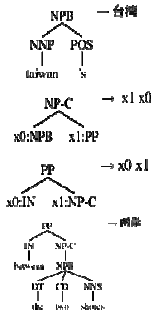


synchronous tree-substitution grammars (STSG)
(Galley et al., 2004; Eisner, 2003)

Extracting syntactic rules

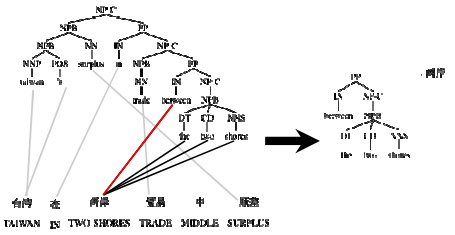


Rules can...



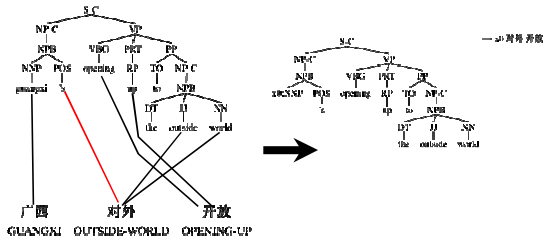
- capture phrasal translation
- reorder parts of the tree
- traverse the tree without reordering
- insert (and delete) words

Bad alignments make bad rules



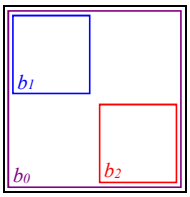
This isn't very good, but let's look at a worse example...

Sometimes they're really bad



One bad link makes a totally unusable rule!

Discriminative Block ITG



Features

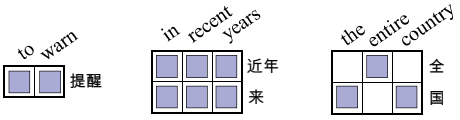
$\varphi(b_0, s, s')$

$\varphi(b_1, s, s')$

$\varphi(b_2, s, s')$

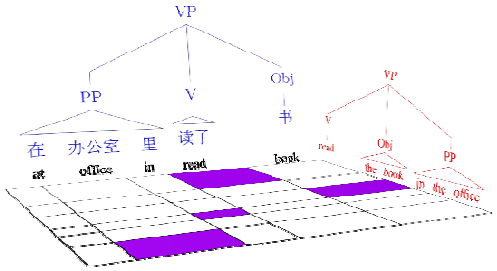
$\phi(\mathcal{A}) = \sum_{b \in \mathcal{A}} \phi(b, s, s')$

$P(\mathcal{A}) \propto \exp\langle \theta, \phi(\mathcal{A}) \rangle$



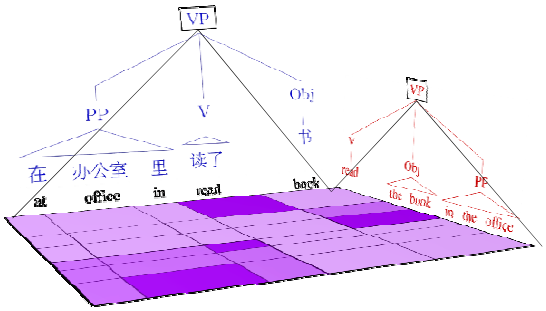
[Haghighi, Blitzer, Denero, and Klein, ACL 09]

Syntactic Correspondence

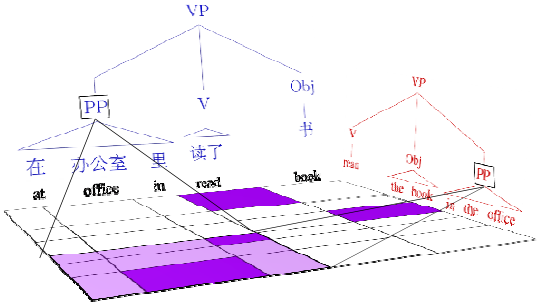


Build a model p_θ (, , 中文, EN)

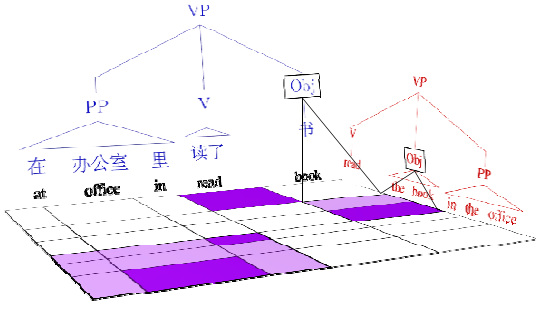
Synchronous Grammars?



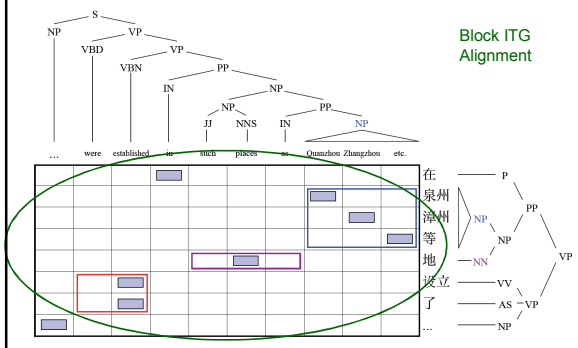
Synchronous Grammars?



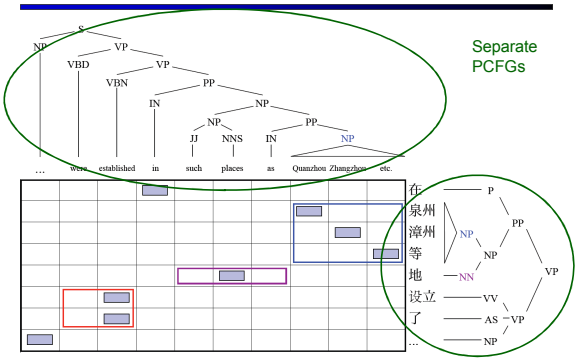
Synchronous Grammars?



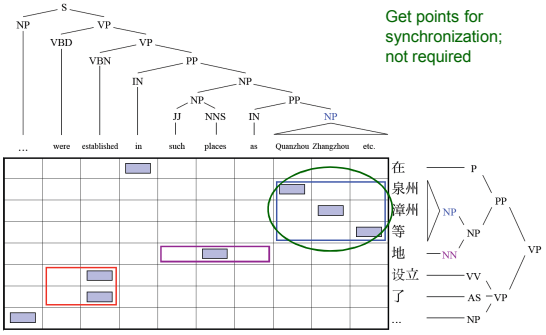
Adding Syntax: Weak Synchronization



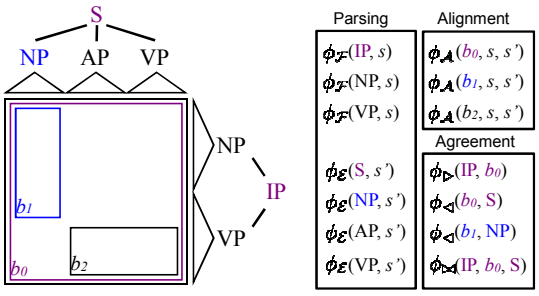
Adding Syntax: Weak Synchronization



Adding Syntax: Weak Synchronization



Weakly Synchronous Features



Weakly Synchronous Model

p_θ (▲, ▲, ▦ | EN, 中文)

Feature Type 1: Word Alignment

ϕ(▦ , EN , 中文)

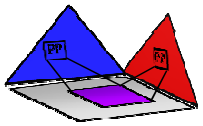
办公室 office [HBDK09]

Feature Type 2: Monolingual Parser

ϕ(▲ , EN)

Feature Type 3: Agreement

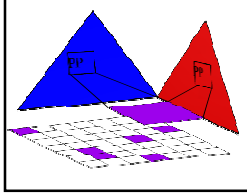
ϕ(▲, ▲, ▦)



Inference: Structured Mean Field

- Problem: Summing over weakly aligned hypotheses is intractable
- Factored approximation: p_θ (▲, ▲, ▦ | EN, 中文) $\approx q(▲)q(▲)q(▦)$
- Set q to minimize KL ($q(▲)q(▲)q(▦)$, p_θ (▲, ▲, ▦ | EN, 中文))

Algorithm

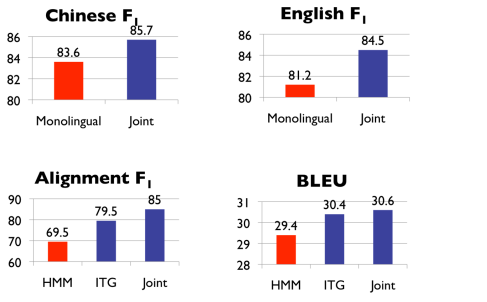


1) Initialize: $q(▲)\ q(▲)\ q(▦)$

2) Iterate:

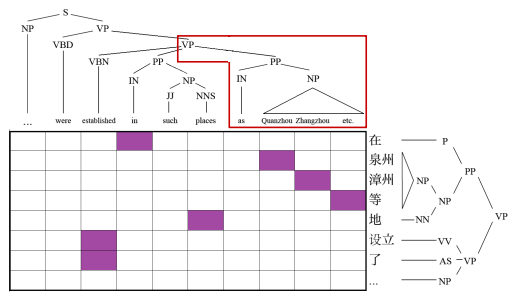
$$q(▲) \propto \exp\{\langle\theta, \phi(▲, E_q(▲), E_q(▦))\rangle\}$$
$$q(▲) \propto \exp\{\langle\theta, \phi(E_q(▲), ▲, E_q(▦))\rangle\}$$
$$q(▦) \propto \exp\{\langle\theta, \phi(E_q(▲), E_q(▲), ▦)\rangle\}$$

Results

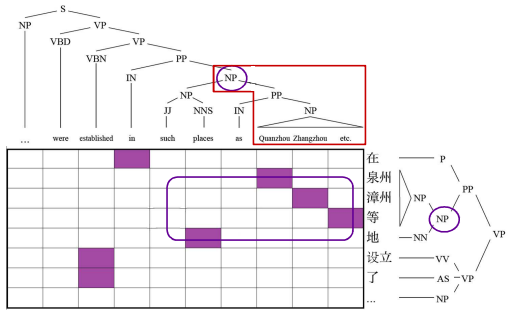


[Burkett, Blitzer, and Klein, NAACL 10]

Incorrect English PP Attachment



Corrected English PP Attachment



Improved Translations

目前	导致	飞机	相撞	的	原因	尚	不	清楚,	当地	民航	部门	将	对此	展开	调查
Cur-rently	cause	plane	crash	DE	reason	yet	not	clear,	local	civil aero-nautics	bureau	will	toward	open	investi-gations

Reference
At this point the cause of the plane collision is still unclear. The local caa will launch an investigation into this .

Baseline (GIZA++)
The cause of planes is still not clear yet, local civil aviation department will investigate this .

Bilingual Adaptation Model
The cause of plane collision remained unclear, local civil aviation departments will launch an investigation .