C SC 620 Advanced Topics in Natural Language Processing

Lecture 24

4/22

Reading List

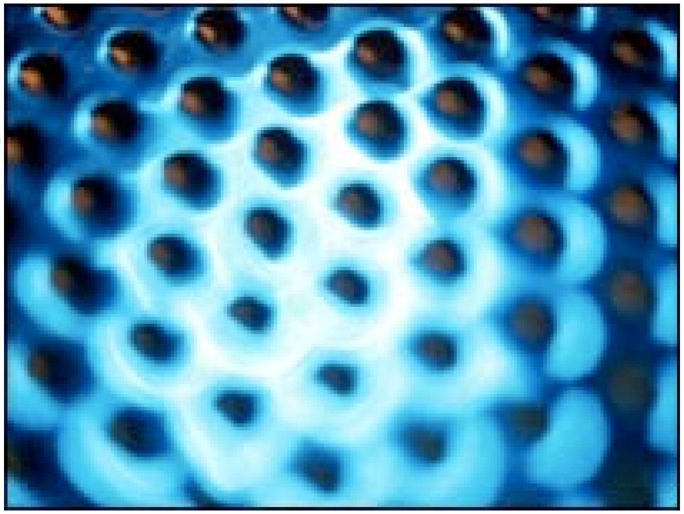
- *Readings in Machine Translation*, Eds. Nirenburg, S. *et al*. MIT Press 2003.
  - 19. Montague Grammar and Machine Translation. Landsbergen, J.
  - 20. Dialogue Translation vs. Text Translation – Interpretation Based Approach. Tsujii, J.-I. And M. Nagao
  - 21. Translation by Structural Correspondences. Kaplan, R. et al.
  - 22. Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation. Boitet, C.
  - 31. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. Nagao, M.
  - **32. A Statistical Approach to Machine Translation. Brown, P. F. et al.**

Language tools for fight on terror

Software to allow security officials to better search and translate documents in foreign languages, especially Arabic, has been demonstrated at a technology show in Las Vegas, as Clark Boyd reports.

There is an old saying in computing - garbage in, garbage out. And never has the world been so awash in digital garbage.

This "needle in a haystack" problem is compounded even further for US intelligence officers on the hunt for, say, Osama Bin Laden.



Hi-tech tools are helping to searching for terror suspects

For starters, American intelligence agencies are short on people who are competent in Arabic, or even want to be.

Natural selection

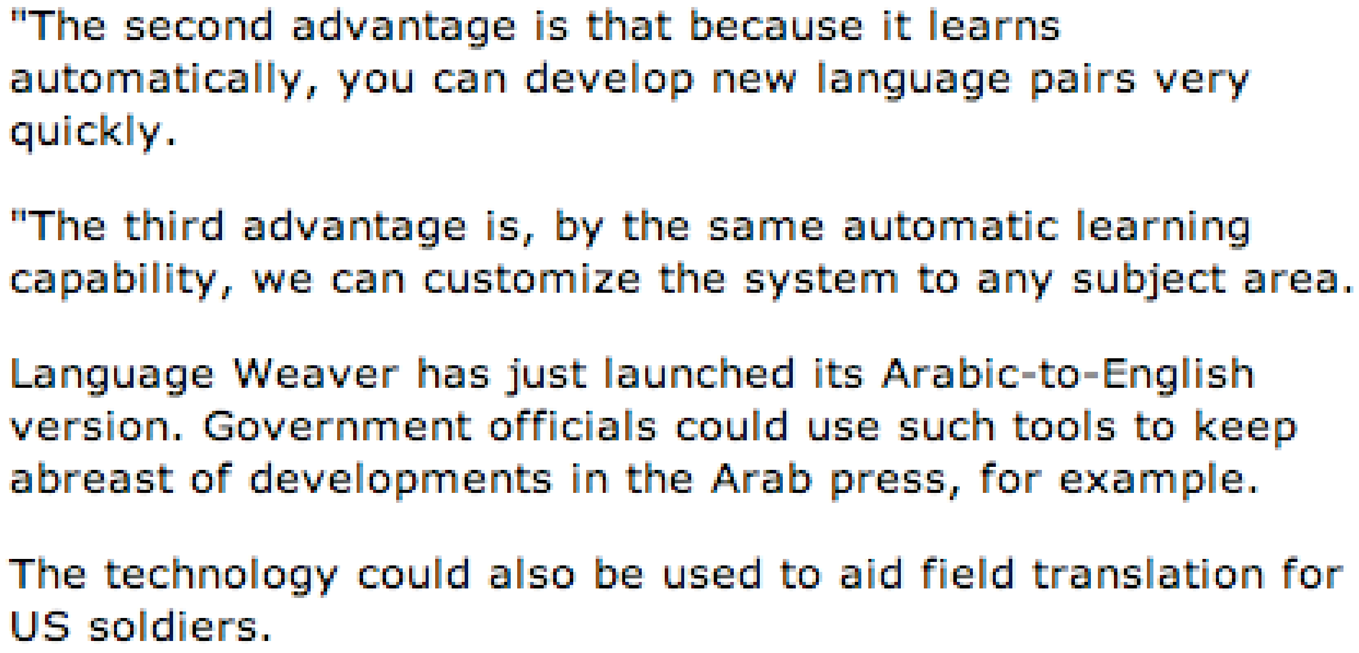
Not all of the language technologies on display in Las Vegas rejected the idea that computers cannot adequately translate Arabic documents directly into English.

Language Weaver is a California-based company that is working with something called statistical natural language processing.

The idea is to train the software using existing human translations. In a sense, the program learns to translate in a more human fashion, the more information is fed to it.

WANTED FUGITIVE
MURDER OF U.S. NATIONALS OUTSIDE AND INSIDE THE UNITED STATES; CONSPIRACY TO MURDER U.S. NATIONALS OUTSIDE THE UNITED STATES; ATTACK AGAINST THE WORLD TRADE CENTER & PENTAGON.
USAMA BIN LADEN
AP
The tools could help in the search for Osama bin Laden

"The first advantage is that it's very natural sounding. The statistical approach gives the system the ability to judge how close it is to real natural language," said Language Weaver's Laurie Gerber.

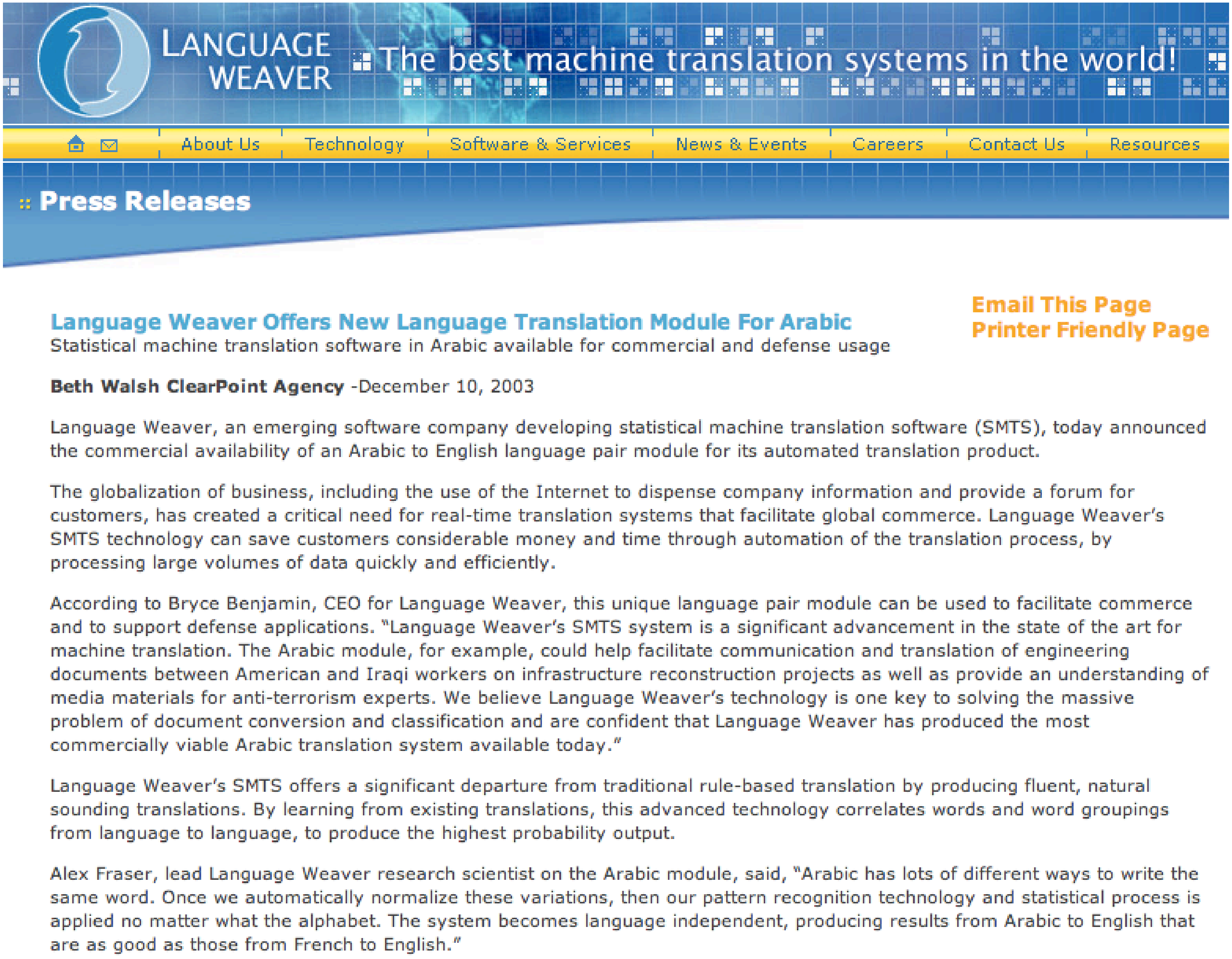


"The second advantage is that because it learns automatically, you can develop new language pairs very quickly.

"The third advantage is, by the same automatic learning capability, we can customize the system to any subject area.

Language Weaver has just launched its Arabic-to-English version. Government officials could use such tools to keep abreast of developments in the Arab press, for example.

The technology could also be used to aid field translation for US soldiers.



:: Press Releases

Language Weaver Offers New Language Translation Module For Arabic

Statistical machine translation software in Arabic available for commercial and defense usage

Beth Walsh ClearPoint Agency -December 10, 2003

Language Weaver, an emerging software company developing statistical machine translation software (SMTS), today announced the commercial availability of an Arabic to English language pair module for its automated translation product.

The globalization of business, including the use of the Internet to dispense company information and provide a forum for customers, has created a critical need for real-time translation systems that facilitate global commerce. Language Weaver's SMTS technology can save customers considerable money and time through automation of the translation process, by processing large volumes of data quickly and efficiently.

According to Bryce Benjamin, CEO for Language Weaver, this unique language pair module can be used to facilitate commerce and to support defense applications. "Language Weaver's SMTS system is a significant advancement in the state of the art for machine translation. The Arabic module, for example, could help facilitate communication and translation of engineering documents between American and Iraqi workers on infrastructure reconstruction projects as well as provide an understanding of media materials for anti-terrorism experts. We believe Language Weaver's technology is one key to solving the massive problem of document conversion and classification and are confident that Language Weaver has produced the most commercially viable Arabic translation system available today."

Language Weaver's SMTS offers a significant departure from traditional rule-based translation by producing fluent, natural sounding translations. By learning from existing translations, this advanced technology correlates words and word groupings from language to language, to produce the highest probability output.

Alex Fraser, lead Language Weaver research scientist on the Arabic module, said, "Arabic has lots of different ways to write the same word. Once we automatically normalize these variations, then our pattern recognition technology and statistical process is applied no matter what the alphabet. The system becomes language independent, producing results from Arabic to English that are as good as those from French to English."

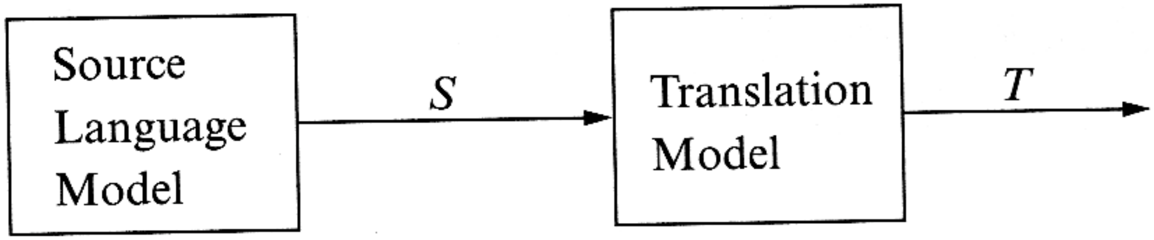
Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Time: Early 1990s
- Emergence of the Statistical Approach to MT and to language modelling in general
  - Statistical learning methods for context-free grammars
    - inside-outside algorithm
- Like the the popular Example-Based Machine Translation (EBMT) framework discussed last time, we avoid the explicit construction of linguistically sophisticated models of grammar
- Why now, and not in the 1950s?
  - Computers 10^5 times faster
  - Gigabytes of storage
  - Large, machine-readable corpora readily available for parameter estimation
  - It’s our turn – symbolic methods have been tried for 40 years

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Machine Translation
  - Source sentence S
  - Target sentence T
  - Every pair (S,T) has a probability
  - P(T|S) = probability target is T given S
  - Bayes' theorem
    - P(S|T) = P(S)P(T|S)/P(T)

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.



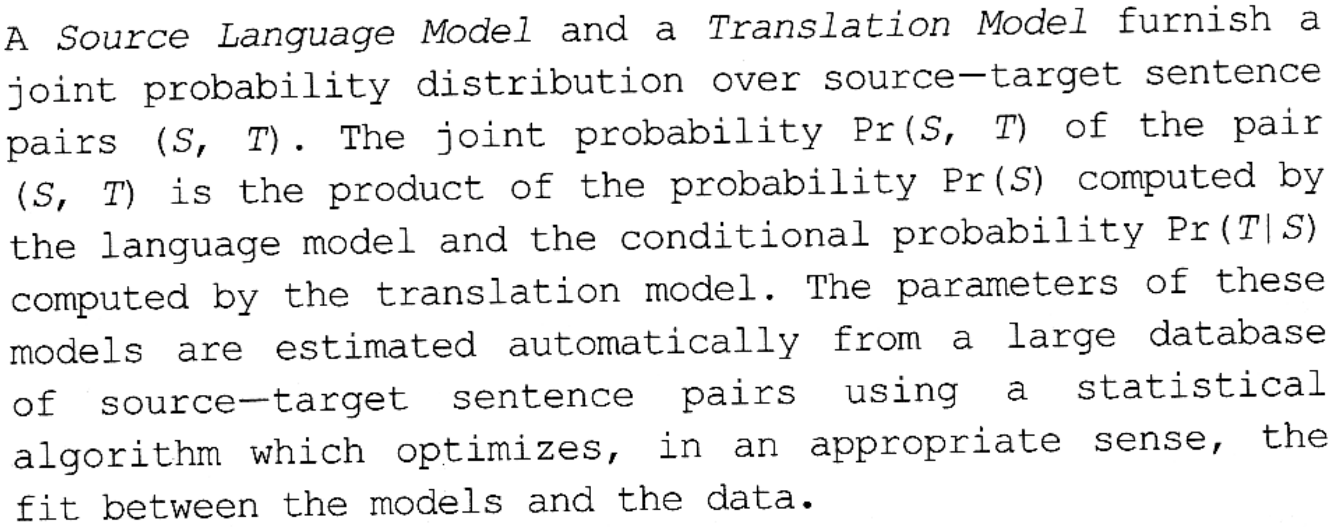
Source Language Model

S

Translation Model

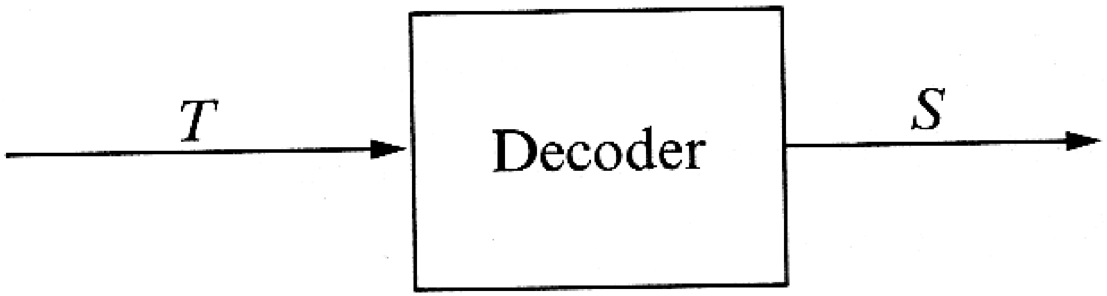
T

$\Pr(S) \times \Pr(T|S) = \Pr(S,T)$

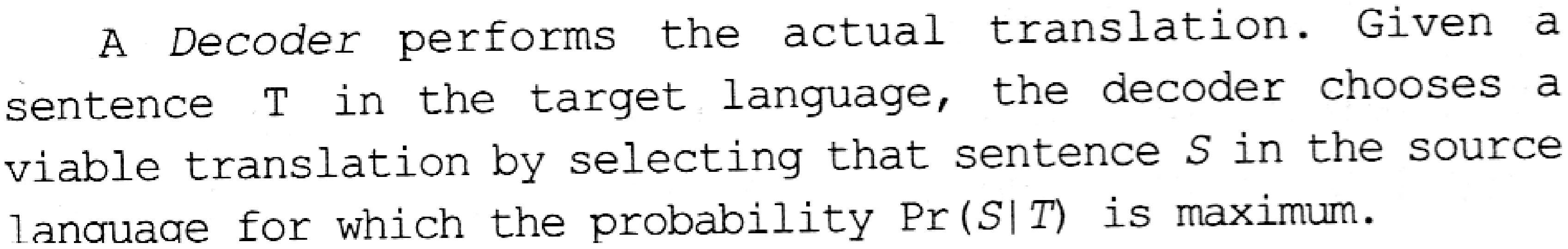


A *Source Language Model* and a *Translation Model* furnish a joint probability distribution over source—target sentence pairs (S, T). The joint probability $\Pr(S, T)$ of the pair (S, T) is the product of the probability $\Pr(S)$ computed by the language model and the conditional probability $\Pr(T|S)$ computed by the translation model. The parameters of these models are estimated automatically from a large database of source—target sentence pairs using a statistical algorithm which optimizes, in an appropriate sense, the fit between the models and the data.

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.



$$S = \underset{S}{\operatorname{argmax}} \Pr(S|T) = \underset{S}{\operatorname{argmax}} \Pr(S,T)$$



A *Decoder* performs the actual translation. Given a sentence T in the target language, the decoder chooses a viable translation by selecting that sentence *S* in the source language for which the probability Pr(*S*|*T*) is maximum.

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- The Language Model: P(S)
  - bigrams:
    - w_1 w_2 w_3 w_4 w_5
    - w_1w_2, w_2w_3, w_3w_4, w_4w_5
  - sequences of words
    - $S = w_1 \ldots w_n$
    - $P(S) = P(w_1)P(w_2| w_1)\ldots P(w_n | w_1 \ldots w_{n-1})$
      - *product of probability of w_i given preceding context for w_i*
    - problem: we need to know too many probabilities
  - bigram approximation
    - limit the context
    - $P(S) \approx P(w_1)P(w_2| w_1)\ldots P(w_n | w_{n-1})$
  - bigram probability estimation from corpora
    - $P(w_i| w_{i-1}) \approx freq(w_{i-1}w_i)/freq(w_{i-1})$ in a corpus

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- The Language Model: P(S)
  - *n-gram models used successfully in speech recognition*
  - could use trigrams:
    - w_1 w_2 w_3 w_4 w_5
    - $w_1w_2w_3$, $w_2w_3w_4$, $w_3w_4w_5$
  - problem
    - need even more data for parameter estimation
    - sparse data problem even with large corpora
    - handled using smoothing
      - interpolate for missing data
      - estimate trigram probabilities from bigram and unigram data

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- The Translation Model: P(T|S)
  - Alignment model:
    - assume there is a transfer relationship between source and target words
    - not necessarily 1-to-1
  - Example
    - S = $w_1\ w_2\ w_3\ w_4\ w_5\ w_6\ w_7$
    - T = $u_1\ u_2\ u_3\ u_4\ u_5\ u_6\ u_7\ u_8\ u_9$
    - w_4 -> $u_3\ u_5$
    - **fertility** of $w_4 = 2$
    - **distortion** w_5 -> u_9

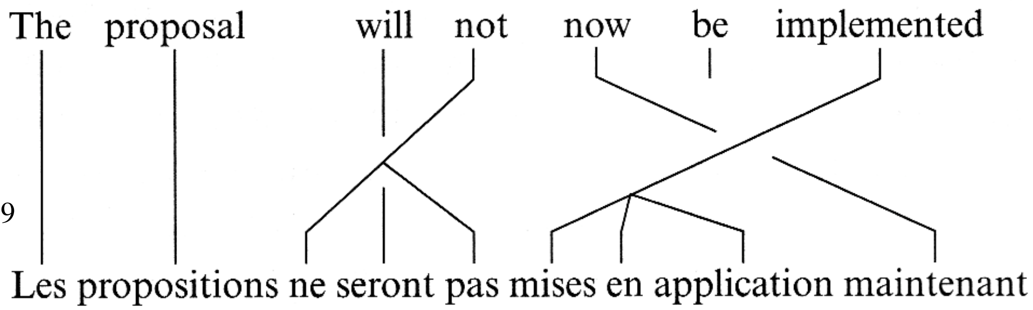

The proposal will not now be implemented
Les propositions ne seront pas mises en application maintenant

Figure 32.3
Alignment example.

Paper 32. A Statistical Approach to Machine Translation.
Brown, P. F. et al.

- Alignment notation
  - *use word positions in parentheses*
  - *no word position, no mapping*
  - Example
    - (Les propositions ne seront pas mises en application maintenant | The(1) proposal(2) will(4) not(3,5) now(9) be implemented(6,7,8))
    - *This particular alignment is not correct, an artifact of their algorithm*

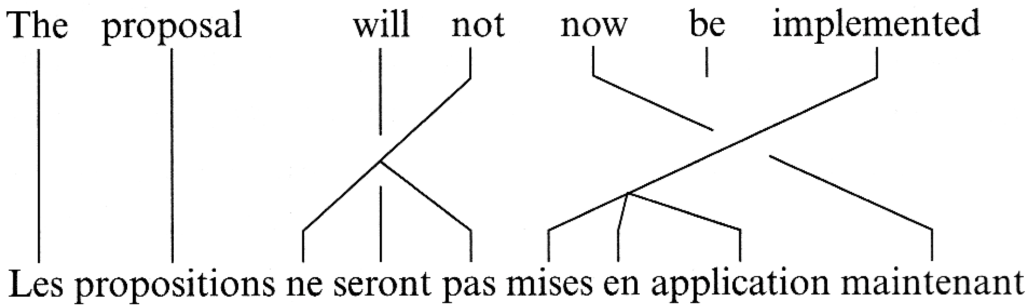

The proposal will not now be implemented
Les propositions ne seront pas mises en application maintenant

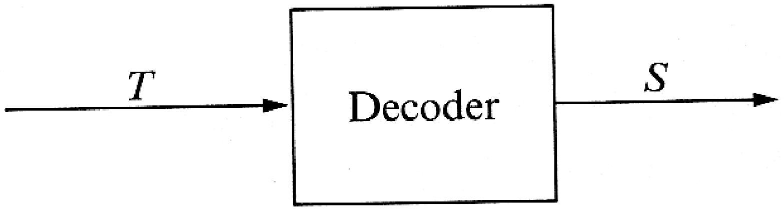
Figure 32.3
Alignment example.

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- How to compute probability of an alignment?
  - Need to estimate
    - Fertility probabilities
      - P(fertility=n|w) = probability word *w* has fertility *n*
    - Distortion probabilities
      - P(i|j,l) = probability target word is at position *i* given source word at position *j* and *l* is the length of the target
  - Example
    - (Le chien est battu par Jean | John(6) does beat(3,4) the(1) dog(2))
      - P(f=1|*John*)P(*Jean*|*John*) x
      - P(f=0|*does*) x
      - P(f=2|*beat*)P(*est*|*beat*)P(*battu*|*beat*) x
      - P(f=1|*the*)P(*Le*|*the*) x
      - P(f=1|*dog*)P(*chien*|*dog*) x
      - P(f=1|<*null*>)P(*par*|<*null*>) x *distortion probabilities…*

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Not done yet
  - Given T
  - translation problem is to find S that maximizes P(S)P(T|S)
  - can't look for all possible S in the language
- Idea (Search):
  - construct best S incrementally
  - start with a highly likely word transfer
  - and find a valid alignment
  - extending candidate S at each step
  - (Jean aime Marie | *)
  - (Jean aime Marie | John(1) *)



$$S = \underset{S}{\text{argmax}}\,\Pr(S|T) = \underset{S}{\text{argmax}}\,\Pr(S,T)$$

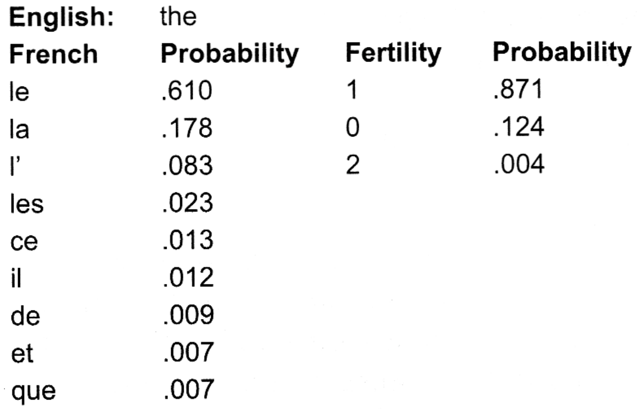
- Failure?
  - best S not a good translation
    - language model failed or
    - translation model failed
  - couldn't find best S
    - search failure

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Parameter Estimation
  - English/French
    - from the Hansard corpus
      - 100 million words
      - bilingual Canadian parliamentary proceedings
      - unaligned corpus
  - Language Model
    - P(S) from bigram model
  - Translation Model
    - how to estimate this with an unaligned corpus?
    - Used EM (Estimation and Maximization) algorithm, an iterative algorithm for re-estimating probabilities
    - Need
      - $P(u|w)$ for words *u* in T and *w* in S
      - $P(n|w)$ for fertility *n* and *w* in S
      - $P(i|j,l)$ for target position *i* and source position *j* and target length *l*

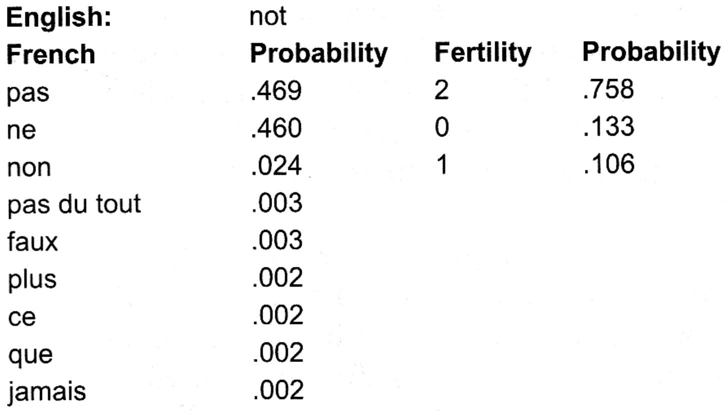
Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Experiment 1: Parameter Estimation for the Translation Model
  - Pick 9,000 most common words for French and English
  - 40,000 sentence pairs
  - 81,000,000 parameters
  - Initial guess: minimal assumptions



English:	the		
French	**Probability**	**Fertility**	**Probability**
le	.610	1	.871
la	.178	0	.124
l'	.083	2	.004
les	.023		
ce	.013		
il	.012		
de	.009		
et	.007		
que	.007		

Figure 32.4
Probabilities for *the*.

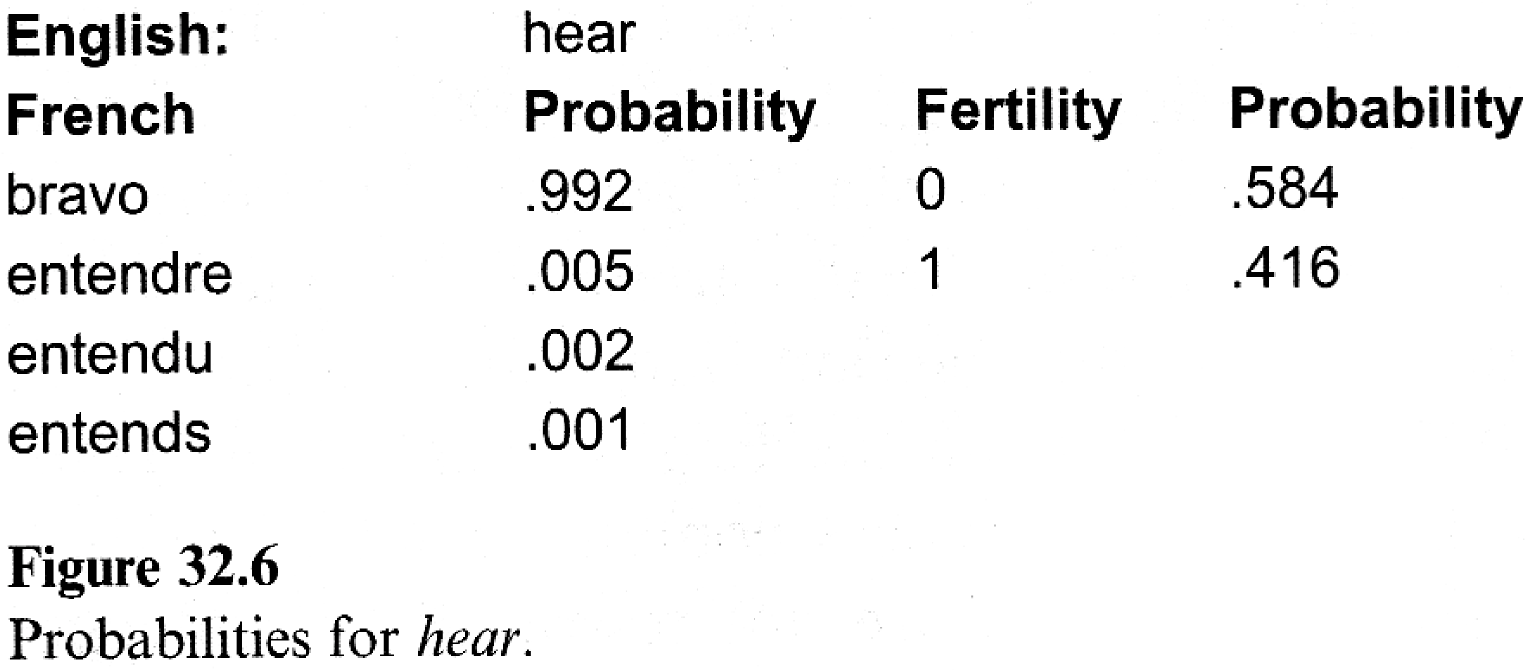


English:	not		
French	**Probability**	**Fertility**	**Probability**
pas	.469	2	.758
ne	.460	0	.133
non	.024	1	.106
pas du tout	.003		
faux	.003		
plus	.002		
ce	.002		
que	.002		
jamais	.002		

Figure 32.5
Probabilities for *not*.

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Experiment 1: results
  - (English) Hear, hear!
  - (French) Bravo!



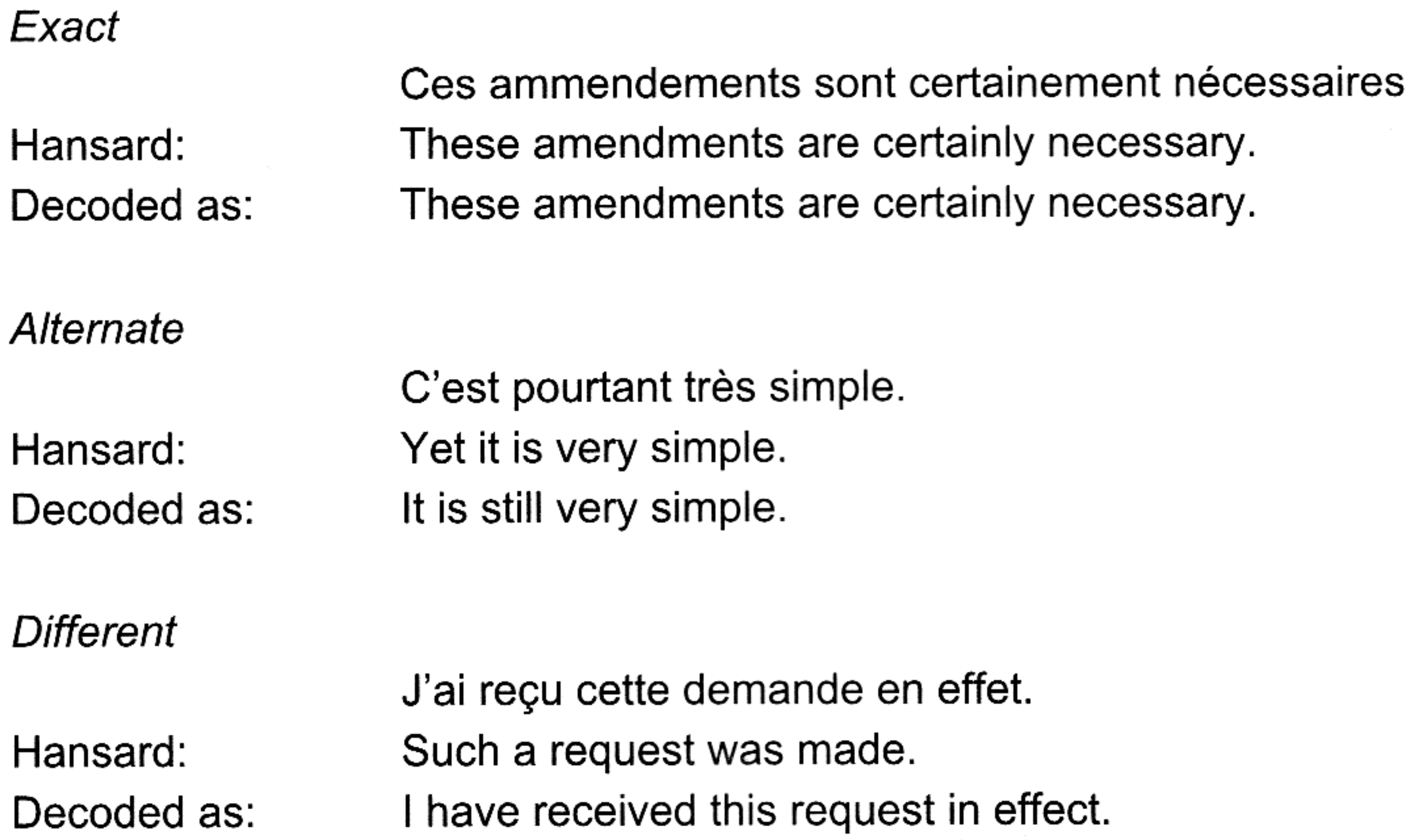
English:	hear		
French	**Probability**	**Fertility**	**Probability**
bravo	.992	0	.584
entendre	.005	1	.416
entendu	.002		
entends	.001		

Figure 32.6
Probabilities for *hear*.

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

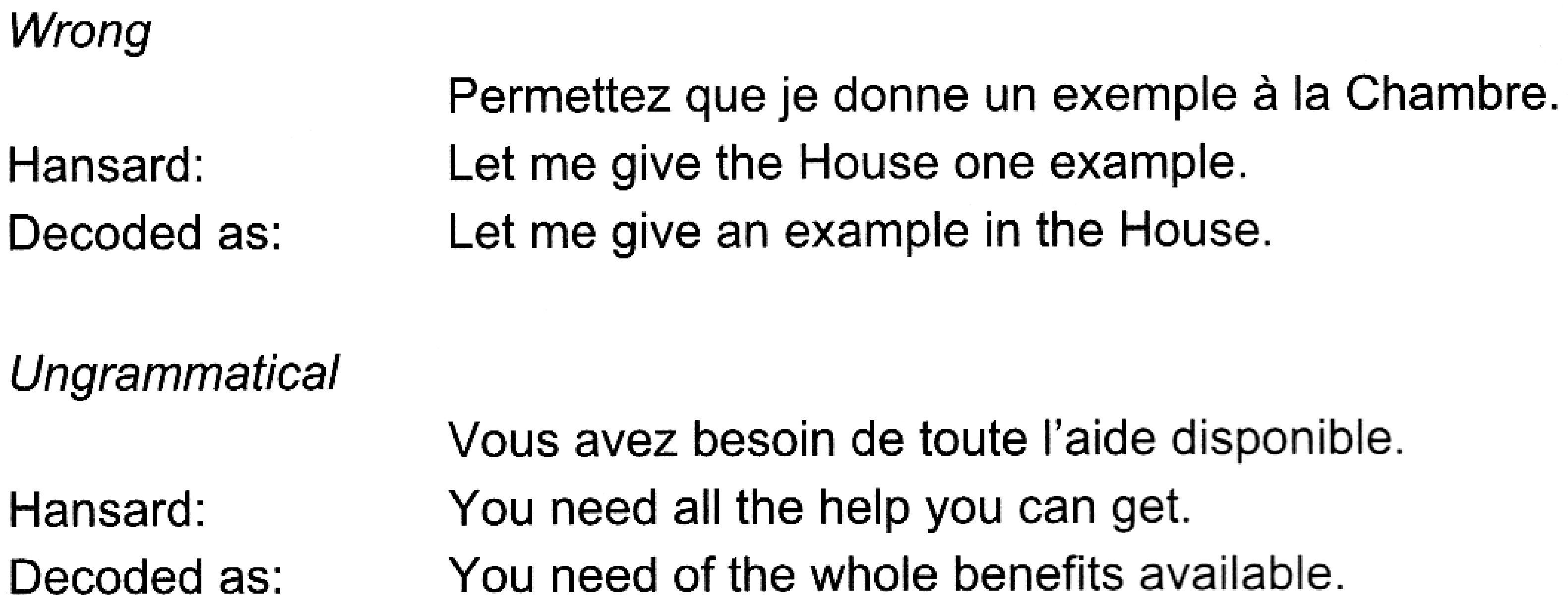
- Experiment 2: Translation from French to English
  - Make task manageable
    - English lexicon
      - 1,000 most frequent English words in corpus
    - French lexicon
      - 1,700 most frequent French words in translations completely covered by the selected English words
    - 117,000 sentence pairs with words covered by the lexicons
    - 17 million parameters estimated for the translation model
    - bigram model of English
      - 570,000 sentences
      - 12 million words
  - 73 test sentences
    - Categories: (exact, alternate, different), wrong, ungrammatical

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.



Exact	
	Ces ammendements sont certainement nécessaires
Hansard:	These amendments are certainly necessary.
Decoded as:	These amendments are certainly necessary.
Alternate	
	C’est pourtant très simple.
Hansard:	Yet it is very simple.
Decoded as:	It is still very simple.
Different	
	J’ai reçu cette demande en effet.
Hansard:	Such a request was made.
Decoded as:	I have received this request in effect.

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.



Wrong

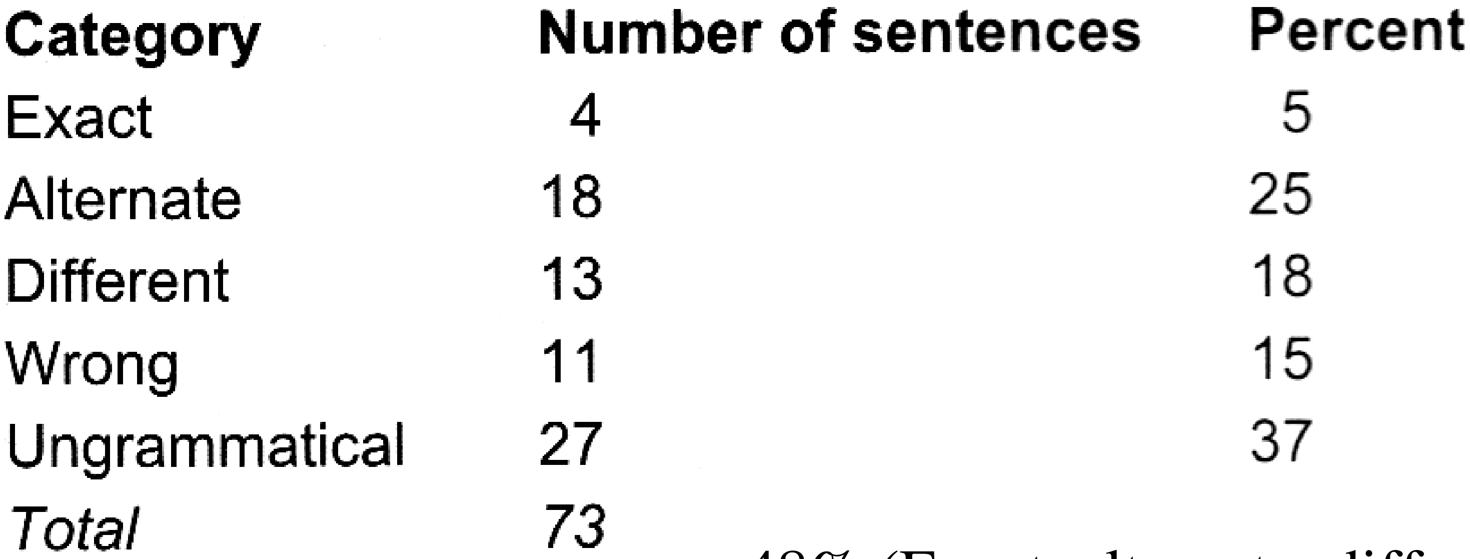
	Permettez que je donne un exemple à la Chambre.
Hansard:	Let me give the House one example.
Decoded as:	Let me give an example in the House.

Ungrammatical

	Vous avez besoin de toute l'aide disponible.
Hansard:	You need all the help you can get.
Decoded as:	You need of the whole benefits available.

Figure 32.7
Translation examples.

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.



Category	Number of sentences	Percent
Exact	4	5
Alternate	18	25
Different	13	18
Wrong	11	15
Ungrammatical	27	37
Total	*73*	

Figure 32.8
Translation results.

48% (Exact, alternate, different)
Editing
776 keystrokes
1,916 Hansard

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Plans
  - Used only a small fraction of the data available
    - Parameters can only get better…
  - Many-to-one problem
    - only one-to-many allowed in current model
    - can’t handle
      - to go -> aller
      - will … be -> seront
  - No model of phrases
    - displacement of phrases

Paper 32. A Statistical Approach to Machine Translation. Brown, P. F. et al.

- Plans
  - Trigram model
    - perplexity = measure of degree of uncertainty in the language model with respect to a corpus
    - Experiment 2: bigram model (78), trigram model (9)
    - trigram model, general English (247)
  - No morphology
    - stemming will help statistics
  - Could define translation between phrases in a probabilistic phrase structure grammar

Administrivia

- Away next week at the University of Geneva
  - work on your projects and papers
  - reachable by email
- Last class
  - Tuesday May 4th