(Statistical) Machine Translation

Christopher Manning
CS224N/Ling237

(includes slides from Yamada, Knight, Brew, etc.)

2004

Translation (human and machine)

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。

Ref: According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959 billion US dollars of foreign capital, including 40.007 billion US dollars of direct investment from foreign businessmen.

IBM4:
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

Yamada/Knight:
today's available data of the Ministry of Foreign Trade and Economic Cooperation shows that china's actual utilization of November this year will include 40.007 billion US dollars for the foreign direct investment among 46.959 billion US dollars in foreign capital

Warren Weaver

- "Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' "
  - Warren Weaver (1955:18, quoting a letter he wrote in 1947)

Machine Translation History

- 1950's: Intensive research activity in MT
- 1960's: Direct word-for-word replacement
- 1966 (ALPAC): NRC Report on MT
  - Conclusion: MT no longer worthy of serious scientific investigation.
- 1966-1975: `Recovery period'
- 1975-1985: Resurgence (Europe, Japan)
- 1985-present: Gradual Resurgence (US)

http://ourworld.compuserve.com/homepages/WJHutchins/MTS-93.htm

What happened between ALPAC and Now?

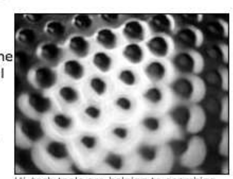
- Need for MT and other NLP applications confirmed
- Change in expectations
- Computers have become faster, more powerful
- WWW
- Political state of the world
- Maturation of Linguistics
- Development of hybrid statistical/symbolic approaches

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.

LANGUAGE WEAVER The best machine translation systems in the world!

About Us | Technology | Software & Services | News & Events | Careers | Contact Us | Resources

Press Releases

Email This Page
Printer Friendly Page

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.

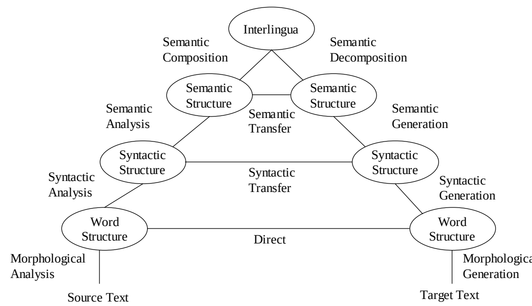
The globalization of business, including the use of the Internet to dispense company information and provide a forum for customers, has created a critical need for real-time translation systems that facilitate global commerce. Language Weaver's SMTS technology can save customers considerable money and 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."

Three MT Approaches: Direct, Transfer, Interlingual (Vauquois triangle)



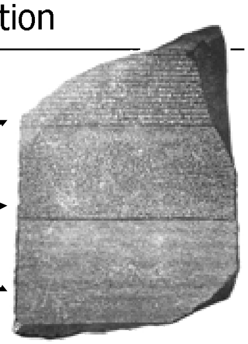
Statistical Solution

- Parallel Texts
  - Rosetta Stone

Hieroglyphs

Demotic

Greek



Statistical Solution

- Parallel Texts
  - Instruction Manuals
  - Hong Kong Legislation
  - Macao Legislation
  - Canadian Parliament Hansards
  - United Nations Reports
  - Official Journal of the European Communities

Text Alignment: Aligning Sentences and Paragraphs

- ***Text alignment*** is useful for bilingual lexicography, MT, but also as a first step to using bilingual corpora for other tasks.
- Text alignment is not trivial because translators do not always translate one sentence in the input into one sentence in the output, although they do so in 90% of the cases.
- Another problem is that of ***crossing dependencies***, where the order of sentences are changed in the translation.

Different Approaches to Text Alignment

- ***Length-Based Approaches***: short sentences will be translated as short sentences and long sentences as long sentences.
- ***Offset Alignment by Signal Processing Techniques***: these approaches do not attempt to align beads of sentences but rather just to align position offsets in the two parallel texts.
- ***Lexical Methods:*** Use lexical information to align beads of sentences.

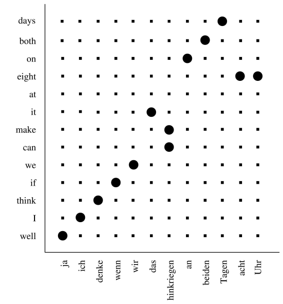
Using Parallel Texts

- Alignment

La Cumbre Iberoamericana de Madrid ha significado un sustancial avance en relación a la celebrada el año anterior en Guadalajara, (México), pues se han aprobado programas concretos, especialmente en el ámbito de la educación y cultura, salud, mundo indígena y medio ambiente.

A Cimeira Ibero-Americana de Madrid constituiu um progresso substancial relativamente à cimeira realizada no ano anterior em Guadalajara, no México, porquanto se aprovaram programas concretos, especialmente no âmbito da educação e da cultura, da saúde, do mundo indígena e do meio ambiente.

Word Alignment Example

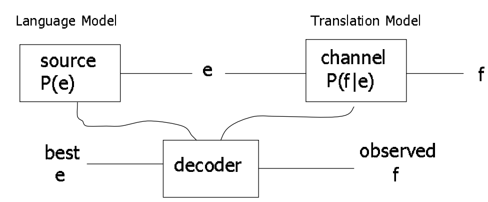


IBM StatMT work

IBM Models 1-5 [Brown, et al., 1993, A Statistical Approach to Machine Translation, *CL*]

- Word-to-word model (no syntax)
- Parameter tables (Model 3)
  - Fertility table: n(k|e)
  - Distortion table: d(tgtPos|srcPos,srcLen,tgtLen)
  - Translation table: t(f|e)
- 2M French-English corpus

Noisy Channel MT Model



$$\operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e P(f|e)P(e)$$

Why not just model P(E|F)?

- It seems circular to model P(E|F) by an approach that requires P(F|E)
- But
  - P(F|E) is the translation specialist
  - P(E) is the English specialist
  - Both can be sloppy

Translation Models

- IBM1 – lexical probabilities only
- IBM2 – lexicon plus absolute position
- HMM – lexicon plus relative position
- IBM3 – plus fertilities
- IBM4 – inverted relative position alignment
- IBM5 – non-deficient version of model 4

[Brown, et.al. 93, Vogel, et.al. 96]

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

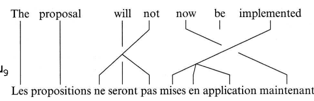
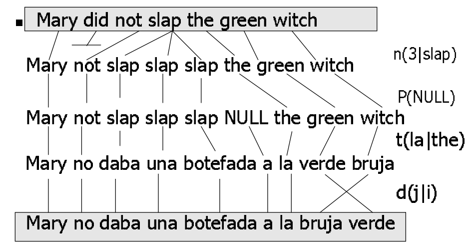


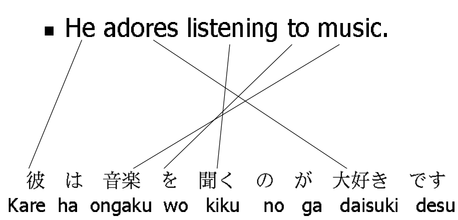
Figure 32.3
Alignment example.

IBM Model 3



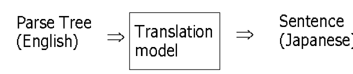
[Al-Onaizan and Knight, 1998]

Yamada and Knight (2001): The need for phrasal syntax



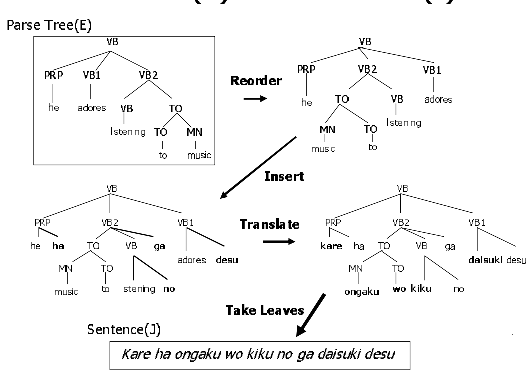
Syntax-based Model

- E→J Translation (Channel) Model

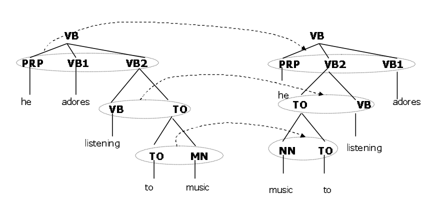


- Preprocess English by a parser
- Probabilistic Operations on a parse-tree
  1. Reorder child nodes
  2. Insert extra nodes
  3. Translate leaf words

Parse Tree(E) → Sentence (J)



1. Reorder



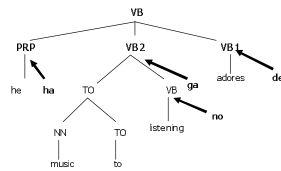
P(PRP VB1 VB2 → PRP VB2 VB1) = 0.723
P(VB TO → TO VB) = 0.749
P(TO NN → NN TO) = 0.893

Conditioning Feature = Child label Sequence

Parameter Table: Reorder

Original Order	Reordering	P(reorder|original)
PRP VB1 VB2	PRP VB1 VB2	0.074
	PRP VB2 VB1	**0.723**
	VB1 PRP VB2	0.061
	VB1 VB2 PRP	0.037
	VB2 PRP VB1	0.083
	VB2 VB1 PRP	0.021
VB TO	VB TO	0.107
	TO VB	**0.893**
TO NN	TO NN	0.251
	NN TO	**0.749**

2. Insert



P(none|TOP-VB) = 0.735

⋮

P(right|VB-PRP) * P(ha) = 0.652 * 0.219

P(right|VB-VB) * P (ga) = 0.252 * 0.062

⋮

P(none|TO-TO) = 0.900

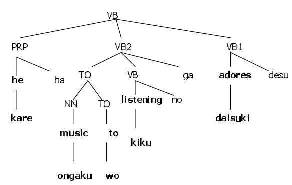
Conditioning Feature = Parent Label & Node Label (position)
none (word selection)

Parameter Table: Insert

Parent label node level	TOP VB	VB VB	VB TO	TO TO	TO NN	TO NN
P (none)	**0.735**	0.687	0.344	**0.700**	**0.900**	**0.800**
P (left)	0.004	0.061	0.004	0.030	0.003	0.096
P (right)	0.260	**0.252**	**0.652**	0.261	0.097	0.104

W	P (insert-w)
ha	**0.219**
ta	0.131
wo	0.099
no	**0.094**
ni	0.080
te	0.078
ga	**0.062**
⋮	⋮
desu	**0.0007**
⋮	⋮

3. Translate



P (he —— kare) = 0.952
P (music —— ongaku) = 0.900
P (to — wo) = 0.038
P (listening — kiku) = 0.333
P (adore — daisuki) = 1.000

Conditioning Feature— word (E) identity

Parameter Table: Translate

E	adores	he	listening	music	to
J	**daisuki 1.000**	**kare 0.952** NULL 0.016 nani 0.005 da 0.003 shi 0.003 ⋮ ⋮	**kiku 0.333** kii 0.333 mi 0.333	**ongaku 0.900** naru 0.100	ni 0.216 NULL 0.204 **to 0.133** no 0.046 wo 0.038 ⋮ ⋮

Note: Translation to NULL = deletion

Automatic Parameter Estimation

- EM Algorithm [Dempster et al., 1977]

--- Iteratively update parameters to seek for maximum likelihood of training data

- Polynomial algorithm

--- Forward-Backward algorithm for HMM [Baum, 1972]

--- Inside-Outside algorithm for PCFG [Baker, 1979] (dynamic programming)

Experiment

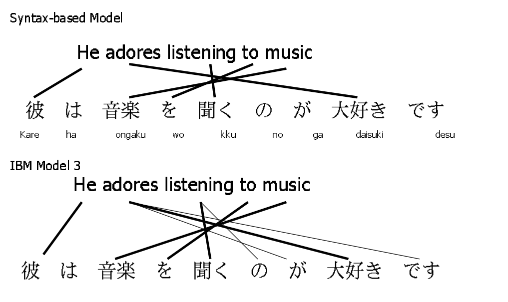
- Training Corpus: J-E 2K sentence pairs
- J: Tokenized by Chasen [Matsumoto, et al., 1999]
- E: Parsed by Collins Parser [Collins, 1999]
  --- Trained: 40K Treebank, Accuracy: ~90%
- E: Flatten parse tree
  --- To Capture word-order difference (SVO->SOV)
- EM Training: 20 Iterations
  --- 50 min/iter (Sparc 200Mhz 1-CPU) or
  --- 30 sec/iter (Pentium3 700Mhz 30-CPU)

Result: Alignments

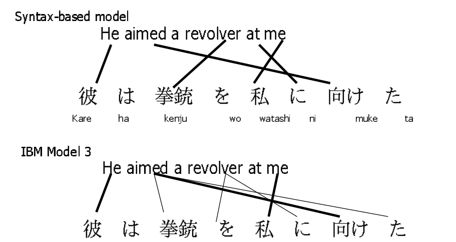
	Ave. Score	# perf sent
Y/K Model	0.582	10
IBM Model 5	0.431	0

- Ave. by 3 humans for 50 sents
- okay(1.0), not sure(0.5), wrong(0.0)
- precision only

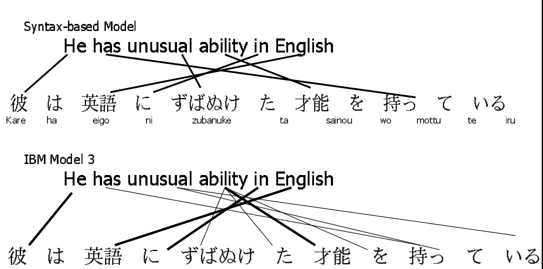
Result: Alignment 1



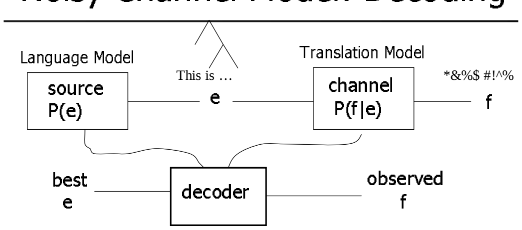
Result: Alignment 2



Result: Alignment 3



Noisy Channel Model: Decoding



$$\arg\max_e P(e|f) = \arg\max_e P(f|e)P(e)$$

Decoding is frequently exponential in Stat MT!

Decoding

- Reverse direction of translation channel
  - English Parse Tree ← foreign sentence
- Use Trigram for LM
- Decoding as parsing
  - - expand English grammar with model operations (reorder,insert,translate)
  - - additional info (cost, reorder)

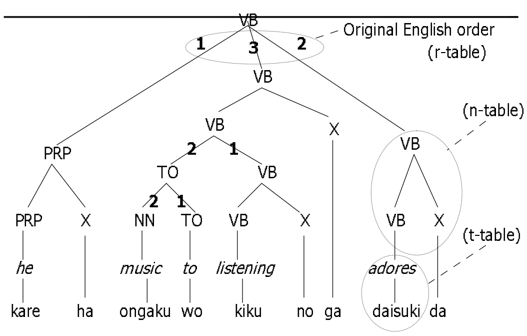
Decoding Grammar

Base English grammar
- VB → PRP VB1 VB2
- VB → VB TO
- VB → TO NN
- PRP → "He"
- VB → "adores"
- VB → "listening"
- ...

r-table / n-table / t-table

Expand by Model
- VB → PRP VB2 VB1
- VB → VB1 VB2 PRP
- ...
- VB → X VB
- VB → VB X
- X → "wo"
- ...
- "He" → "kare"
- "adores" → "daisuki"
- ...

Decoded Tree



Decoding Grammar for C/E 3M corpus

Base English grammar
- VB → PRP VB1 VB2
- VB → VB TO
- VB → TO NN
- ...

18K rules

- PRP → "He"
- VB → "adores"
- VB → "listening"
- ...

65K Ewords
14K Cwords

r-table / t-table

Expand by Model
- VB → PRP VB2 VB1
- VB → VB1 VB2 PRP
- ...

141K rules

- VB → X VB
- VB → VB X
- X → "wo"
- ...
- "He" → "kare"
- "adores" → "daisuki"
- ...

4M rules

Reducing Decoder search space

- Beam search
  - Dynamic-programming parser
  - Bottom-up within beam-width (similar to [Collins 1999])
- Prune decoding grammar
  - prune by rule likelihood
  - Use extra statistics outside of model

Machine Translation

- Usable Technologies
  - "Translation memories" to aid translator
  - Low quality screening/web translators
- Technologies
  - Traditional: Systran (Altavista Babelfish, Google) is now seen as a limited success
  - Statistical MT over huge training sets is quite successful (LanguageWeaver, Microsoft, Google's future?)
- Key ideas for the future
  - Statistical phrases
  - Syntax based models
  - Better language models in other respects (e.g., grammar)
  - Usably efficient decoding models (by restricting model?)