

Machine Translation Systems

Christopher Manning
CS224N / Ling 284

[Based on slides by Kevin Knight, Dan Klein,
Dan Jurafsky]

MT Evaluation

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")
 - 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)

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

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision
- Note: only works at corpus level (zeros kill it); there's a smoothed variant for sentence-level

BLEU in Action

枪手被警方击毙。

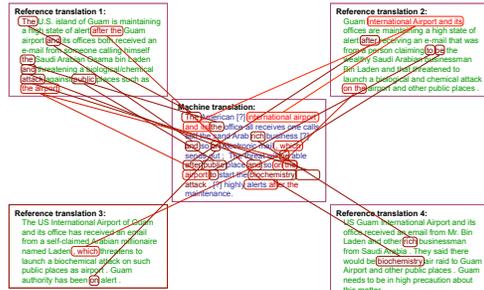
(Foreign Original)

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

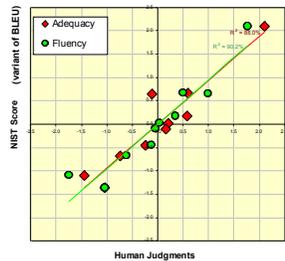
- the gunman was police kill . #1
- wounded police jaya of #2
- the gunman was shot dead by the police . #3
- the gunman arrested by police kill . #4
- the gunmen were killed . #5
- the gunman was shot to death by the police . #6
- gunmen were killed by police ?SUB>0 ?SUB>0 al by the police . #7
- the 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)

Quiz question!

MT Hypothesis: *the gunman was shot dead by police .*

- Ref 1: The gunman was shot to death by the police .
- Ref 2: The cops shot the gunman dead .

• What is the:

- Unigram precision?
- Trigram precision?

Note: punctuation tokens *are* counted in calculation but not sentence boundary tokens

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 really requires *some* automatic metric to allow a rapid development and evaluation cycle.

A complete translation system

Decoding for IBM Models

- Of all conceivable English word strings, find the one maximizing $P(e) \times P(f | e)$
- Decoding is NP hard
 - (Knight, 1999)
- Several search strategies are available
 - Usually a beam search where we keep multiple stacks for candidates covering the same number of source words
- Each potential English output is called a *hypothesis*.

Search for Best Translation

voulez – vous vous taire !

Search for Best Translation

voulez – vous vous taire !
 \ / \ / \ / \ / \ / \ /
 you – you you quiet !

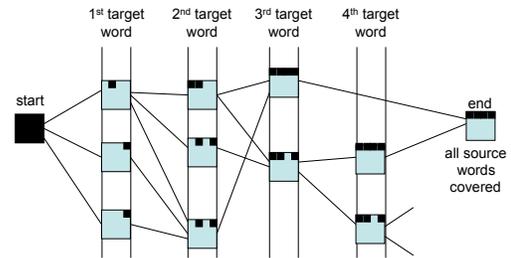
Search for Best Translation

voulez – vous vous taire !
 \ / \ / \ / \ / \ / \ /
 quiet you – you you !

Search for Best Translation

voulez – vous vous taire !
 \ / \ / \ / \ / \ / \ /
 you shut up !

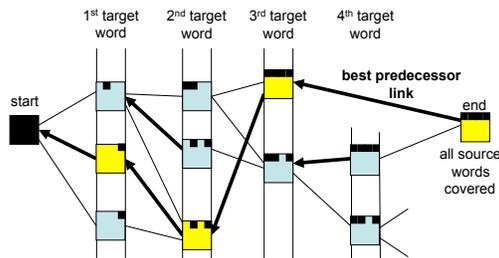
Dynamic Programming Beam Search



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]

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

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

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

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

What StatMT people do in the privacy of their own homes

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

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

$$\operatorname{argmax}_e P(e)^{1.9} \times P(f | e) \quad \dots \text{ works better!}$$

Which model are you now paying more attention to?

What StatMT people do in the privacy of their own homes

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

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

$$\operatorname{argmax}_e P(e)^{1.9} \times P(f | e) \times 1.1^{\text{length}(e)}$$

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

What StatMT people do in the privacy of their own homes

$$\operatorname{argmax}_e P(e)^{1.9} \times P(f | e) \times 1.1^{\text{length}(e)} \times \text{KS}^{3.7} \dots$$

Lots of knowledge sources vote on any given hypothesis.
"Knowledge source" = "feature function" = "score component".
Feature function simply scores a hypothesis with a real value.
(May be binary, as in "e has a verb").

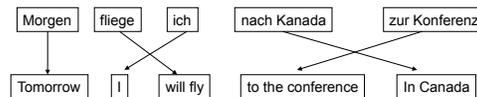
Problem: How to set the weights?
(We look at one way later: maxent models.)

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"
- 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

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

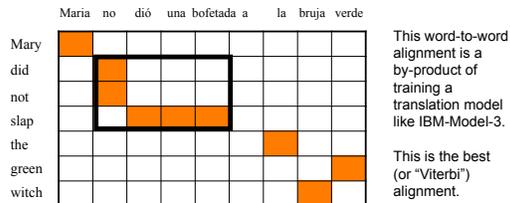
How to Learn the Phrase Translation Table?

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



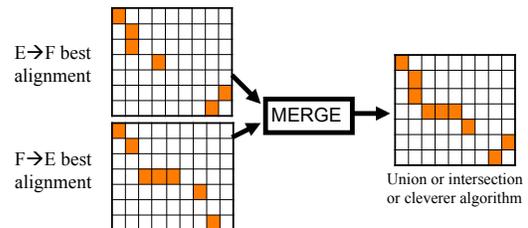
How to Learn the Phrase Translation Table?

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



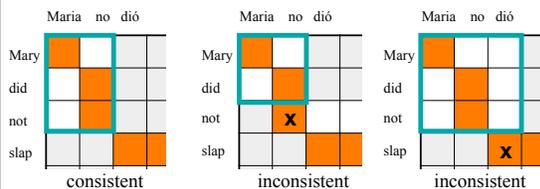
IBM Models are 1-to-Many

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



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*

Phrase Pair Probabilities

- A certain phrase pair (f-f-f, e-e-e) may appear many times across the bilingual corpus.
- No EM training
- Just relative frequency:

$$P(f-f-f | e-e-e) = \frac{\text{count}(f-f-f, e-e-e)}{\text{count}(e-e-e)}$$

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	.			
a	7	people	included	by	france	and	the	russian	the	international	astronautical	of	rapporteur	.	
this	7	and	including	the	from	the	franch	and	the	russian	the	the	the	the	
those	7	among	including	from	the	of	the	russian	of	the	space	members	.		
that	7	persons	including	from	the	of	france	and	to	russian	of	the	astronauts	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				of	astronauts	.	.	
7	deportes	included	come	from	france	and	russia			in	astronautical	personal	.		
7	philtrum	included	those	from	france	and	russia			a	space	astronaut	member	.	
		include	representatives	from	france	and	russia					astronaut		.	
		include	came	from	france	and	russia			by	components			.	
		include	representatives	from	france	and	russia				components			.	
		include	came	from	france	and	russia				components			.	
		includes	coming	from	france	and	russia				components			.	
			coming	from	france	and	russia			's	astronautical	member	.		
			france	and	russian						astronauts			.	
			france	and	russia						astronauts			.	
			and	russia							special	rapporteur		.	
			and	russia							rapporteur			.	
			and	russia							rapporteur			.	
			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	.			
a	7	people	included	by	france	and	the	russian	the	international	astronautical	of	rapporteur	.	
this	7	and	including	the	from	the	franch	and	the	russian	the	the	the	the	
those	7	among	including	from	the	of	the	russian	of	the	space	members	.		
that	7	persons	including	from	the	of	france	and	to	russian	of	the	astronauts	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				of	astronauts	.	.	
7	deportes	included	come	from	france	and	russia			in	astronautical	personal	.		
7	philtrum	included	those	from	france	and	russia			a	space	astronaut	member	.	
		include	representatives	from	france	and	russia					astronaut		.	
		include	came	from	france	and	russia			by	components			.	
		include	representatives	from	france	and	russia				components			.	
		include	came	from	france	and	russia				components			.	
		includes	coming	from	france	and	russia				components			.	
			coming	from	france	and	russia			's	astronautical	member	.		
			france	and	russian						astronauts			.	
			france	and	russia						astronauts			.	
			and	russia							special	rapporteur		.	
			and	russia							rapporteur			.	
			and	russia							rapporteur			.	
			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 航 员 .

the	7	people	including	by	some	and	the	russian	the	the	astronauts	.			
a	7	people	included	by	france	and	the	russian	the	international	astronautical	of	rapporteur	.	
this	7	and	including	the	from	the	franch	and	the	russian	the	the	the	the	
those	7	among	including	from	the	of	the	russian	of	the	space	members	.		
that	7	persons	including	from	the	of	france	and	to	russian	of	the	astronauts	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				of	astronauts	.	.	
7	deportes	included	come	from	france	and	russia			in	astronautical	personal	.		
7	philtrum	included	those	from	france	and	russia			a	space	astronaut	member	.	
		include	representatives	from	france	and	russia					astronaut		.	
		include	came	from	france	and	russia			by	components			.	
		include	representatives	from	france	and	russia				components			.	
		include	came	from	france	and	russia				components			.	
		includes	coming	from	france	and	russia				components			.	
			coming	from	france	and	russia			's	astronautical	member	.		
			france	and	russian						astronauts			.	
			france	and	russia						astronauts			.	
			and	russia							special	rapporteur		.	
			and	russia							rapporteur			.	
			and	russia							rapporteur			.	
			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.

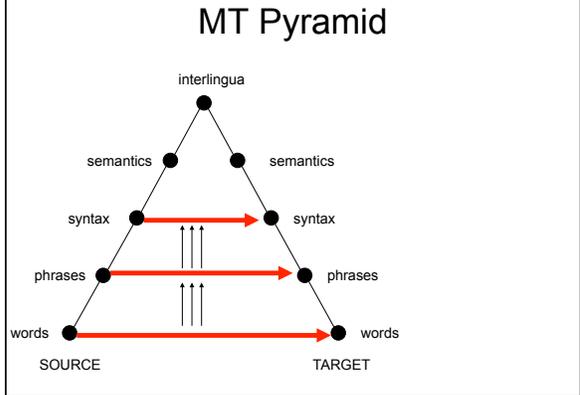
Phrase-Based Translation

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	.			
a	7	people	included	by	france	and	the	russian	the	international	astronautical	of	rapporteur	.	
this	7	and	including	the	from	the	franch	and	the	russian	the	the	the	the	
those	7	among	including	from	the	of	the	russian	of	the	space	members	.		
that	7	persons	including	from	the	of	france	and	to	russian	of	the	astronauts	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				of	astronauts	.	.	
7	deportes	included	come	from	france	and	russia			in	astronautical	personal	.		
7	philtrum	included	those	from	france	and	russia			a	space	astronaut	member	.	
		include	representatives	from	france	and	russia					astronaut		.	
		include	came	from	france	and	russia			by	components			.	
		include	representatives	from	france	and	russia				components			.	
		include	came	from	france	and	russia				components			.	
		includes	coming	from	france	and	russia				components			.	
			coming	from	france	and	russia			's	astronautical	member	.		
			france	and	russian						astronauts			.	
			france	and	russia						astronauts			.	
			and	russia							special	rapporteur		.	
			and	russia							rapporteur			.	
			and	russia							rapporteur			.	
			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.

Syntax and Semantics in Statistical MT



Why Syntax?

- Need much more grammatical output
- Need accurate control over re-ordering
- Need accurate insertion of function words
- Word translations need to depend on grammatically-related words

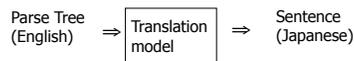
Yamada and Knight (2001): The need for phrasal syntax

- He adores listening to music.

彼は音楽を聞くのが大好きです
Kare ha ongaku wo kiku no ga daisuki desu

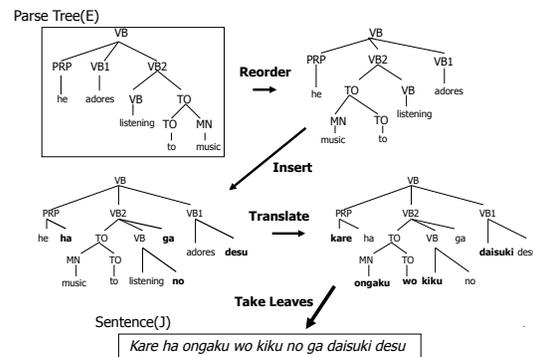
Syntax-based Model

- 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)



Experiment

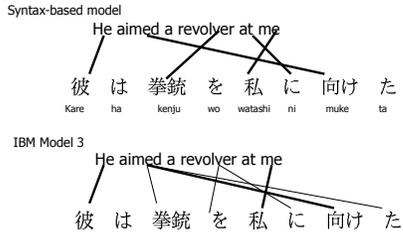
- Training Corpus: J-E 2K sentence pairs
- J: Tokenized by Chasen [Matsumoto, et al., 1999]
- E: Parsed by Collins Parser [Collins, 1999]
 - Trained: 40K Treebank, Accuracy: ~90%
- E: Flatten parse tree
 - To Capture word-order difference (SVO→SOV)
- EM Training: 20 Iterations
 - 50 min/iter (Sparc 200Mhz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

Result: Alignments

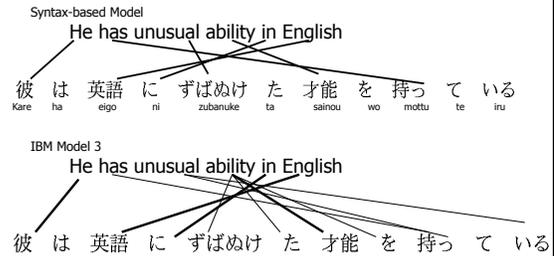
	Ave. Score	# perf sent
Y/K Model	0.582	10
IBM Model 5	0.431	0

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

Result: Alignment 2



Result: Alignment 3



MT Applications

Christopher Manning
CS 224N
2009

MT: The early history (1950s)

- Earliest
- Foundational
- First
- MT
- Little
- sem
- Pro



MT Applications: 1. Traditional

- Traditional scenario:
 - Documents had to be translated for your company/organization. Document production for organization
 - Generally, the quality/accuracy demands are high
 - High cost
 - Though most of it is now done as outsourced piecemeal
- MT tends to be ineffective: The cost of post-translation error correction is too high
- Main technology in the game: translation memory/translation workbench/terminology management
 - E.g., TRADOS.
 - Very slowly, MT technology is starting to be incorporated, but most of the action is in terminology lexicon management

ProZ.com is home to the world's largest community of professional translators

More translators and interpreters: English to Japanese - Portuguese to Japanese - Japanese to Portuguese - Japanese to Portuguese - [New language pairs](#)

T. I. E. Japan Co., Ltd - T. I. E. Japan Co., Ltd | Contact: [CV/Resume](#)

Accuracy, Reliability, Reasonable prices.

Minato-ku, Tokyo, Japan / Native in: Japanese

Profile | **Company profile** | **Contact**

The person shown here is a basic user of this site, he or she may be contacted directly for language-related services.

Working languages: English to Japanese, Portuguese to Japanese, Japanese to Portuguese

Local time: 23:50:27 (GMT+9)

Services: Freelancer and outsource
Translation, Website localization, Software localization, Desktop publishing

Expertise: Business/Commerce (general), Mathematics & Statistics, Mechanics / Mech Engineering, Journalism, IT (Information Technology), Internet, e-Commerce, Electronics / Elect Eng, Telecom(munications), Computers (general), Computers: Systems, Networks

Rates: English to Japanese - Rates: 0.10 - 0.06 USD per word / 40 - 20 USD per hour
Portuguese to Japanese - Rates: 0.13 - 0.08 USD per word / 45 - 25 USD per hour
Japanese to Portuguese - Rates: 0.13 - 0.08 USD per word / 45 - 25 USD per hour
Spanish to Japanese - Rates: 0.13 - 0.08 USD per word / 45 - 25 USD per hour

KudoZ activity: Questions answered: 2, Questions asked: 0 Easy / 0 PRO

Translation education: Univ. de Sao Paulo

Experience: Years of translation experience: 14. Registered at ProZ.com: Oct 2006.

ProZ.com Certified PRO certificate(s): N/A

Credentials: N/A

Memberships: N/A

Software: Adobe Acrobat, Adobe Photoshop, Microsoft Excel, Microsoft Word, PageMaker, Powerpoint

Website: <http://www.tie-net.com>

CV/Resume: [CV/Resume](#)

About me: Languages - Japanese / Brazilian Portuguese / Spanish / English
English-Japanese
Brazilian Portuguese-Japanese, Japanese-Brazilian Portuguese
English-Brazilian Portuguese
Spanish-Japanese
Mother tongue: Japanese

Bad TRADOS Screenshot...

Trados is relatively pricey (high hundreds for PC versions, thousands for server version); seen as necessary productivity tool (Photoshop for translators)

The screenshot shows a window titled 'TRADOS Translator Workbench - TM'. The main text area contains the headline 'FORD ANNOUNCES SAFETY RECALL' and a sub-headline 'FORD ОБЪЯВЛЯЕТ ОБ ОТЗЫВЕ АВТОМОБИЛЕЙ ДУО ОБЕСПЕЧ'. Below this, there is a paragraph in Russian: 'DEARBORN, Mich., Sept. 7 - Ford Motor Company announced today that it is conducting a voluntary safety recall to correct a system interaction that could cause the speed control deactivation switch to over...'.

MT Applications: 2. Web

- Web applications:
 - Dominant scenario: User-initiated translation
 - Crucial difference: The quality doesn't have to be great. The user is usually okay if they can understand the gist of what is going on
- Second scenario
 - Somehow on the web people will accept medium quality results. Accessible information is better than no information
- MT is saved!!! "It's the web, stupid."
 - (But is there money in it?)

AltaVista BabelFish

1997:
Free, automatic translation for the masses.
Revolutionary.

But, what was the underlying technology?
SYSTRAN.

MacOS Dashboard?
SYSTRAN
Google until 2006?
SYSTRAN

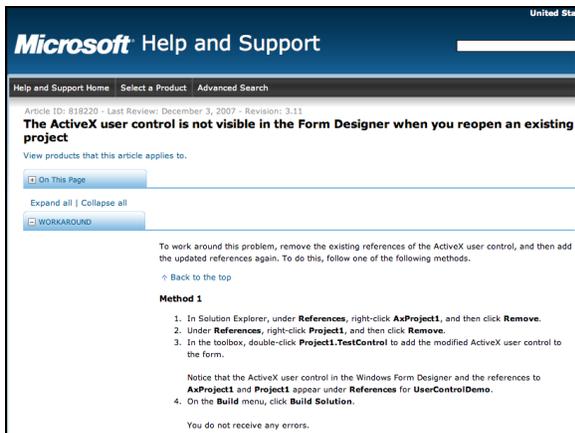
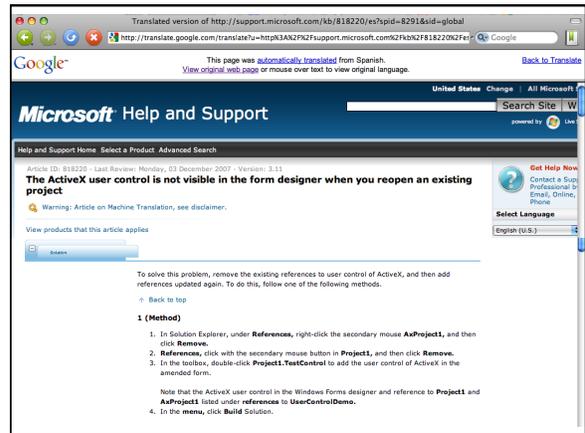
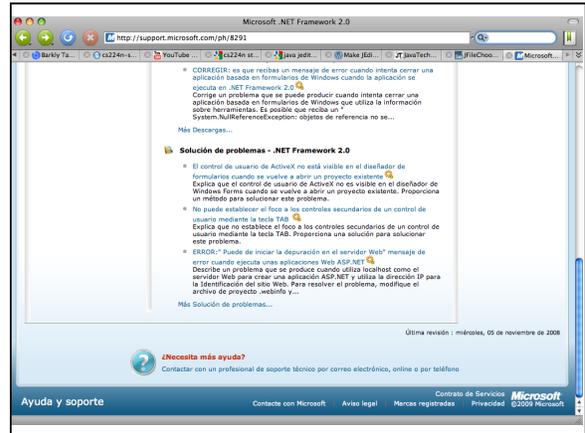
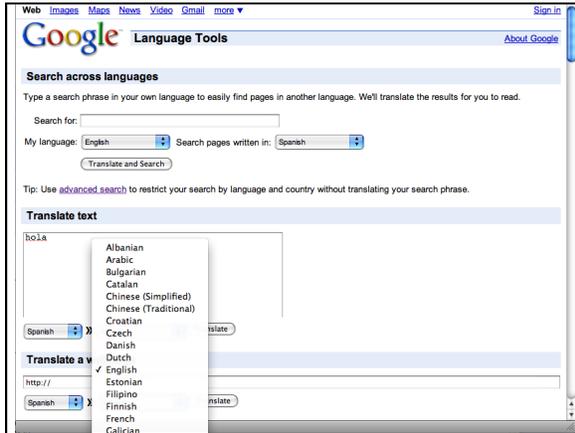
The screenshot shows a news article titled 'DIGITAL ANNOUNCES FIRST INTERNET TRANSLATION CAPABILITY'. The article mentions that Digital Equipment Corporation has broken the internet language barrier with its first European language translation service. It also notes that the service is available for free on the non-English speaking Web surface and can translate information on the internet from English to Web in five. The article is dated May 9, 1997.

The screenshot shows the Wikipedia article for SYSTRAN. The article text includes: 'SYSTRAN, founded by Dr. Peter Toma in 1968^[1], is one of the oldest machine translation companies. SYSTRAN has done extensive work for the United States Department of Defense and the European Commission. SYSTRAN provides the technology for Yahoo! and AltaVista's (Babel Fish) among others, but use of it was ended (circa 2007) for all of the language combinations offered by Google's language toolset. SYSTRAN is used by the Dashboard Translation widget in Mac OS X. Commercial versions of SYSTRAN operate with operating systems Microsoft Windows (including Windows Mobile), Linux and Solaris.'

The screenshot shows the SYSTRAN website homepage. It features a navigation menu with 'PROCESSES', 'SOLUTIONS', 'SUPPORT', 'ABOUT SYSTRAN', 'CONTACT', and 'PARTNERS'. Below the navigation, there is a section for 'SYSTRAN Translation Software' with a list of products: Home & Home Office, Small Business, and Enterprise Solutions. There is also a section for 'An International Benchmark in Translation Software' and 'What is automatic translation?'.

The screenshot shows the LANGUAGE WEAVER website. It features a section for 'LW ENTERPRISE TRANSLATION SERVER' with a diagram of the 'RUN TIME TRANSLATION PROCESS'. The diagram shows an 'ORIGINAL DOCUMENT' being processed by a 'LANGUAGE WEAVER DECODER' to produce a 'TRANSLATED DOCUMENT'. Below the diagram, there are 'SYSTEM REQUIREMENTS' for Client Server 32-bit Platform and Client Server 64-bit Platform. There are also two tables comparing SYSTRAN to other translation services.

TRANSLATION PARAMETERS (Bilingual Data)		LANGUAGE MODEL (Monolingual Data)		
Magnum	Best	19%	The best companies in the globe	27%
	SYSTRAN	17%	World's best companies	11%
	Other	2.3%	Better companies in the world	6.1%
	Trados	2%	The world's companies are best	6.8%
	Other	16.7%		
	Other	1.7%		
	Other	1.1%		
	Other	2.0%		



Machine Translation Summary

- Usable Technologies
 - "Translation memories" to aid translator
 - Low quality screening/web translators
- Technologies
 - Traditional: Systran (Attavista Babelfish, what you got till mid-2006 on Google) is now seen as a limited success
 - Statistical MT over huge training sets is successful (ISI/LanguageWeaver, Microsoft, Google)
- Key ideas of the present/future
 - Statistical phrase based models
 - Syntax based models
 - Better language models (e.g., bigger, using grammar)
 - Better decoding models (e.g., by restricting model?)