Statistical Machine Translation

May 13th, 2014

UNIVERSITÄT DES SAARLANDES

Josef van Genabith
DFKI GmbH
Josef.van_Genabith@dfki.de

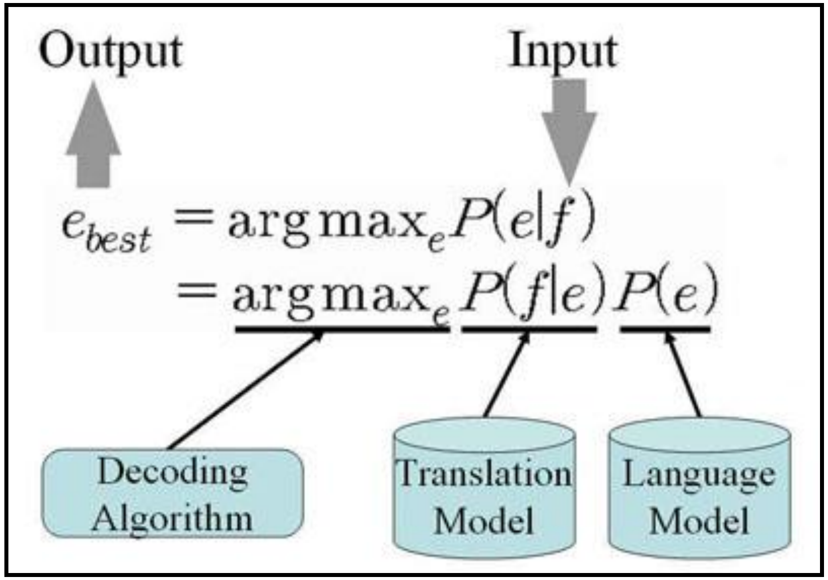
Language Technology II
SS 2014

With some additional slides from Chris Dyer
MT Marathon 2011 and Sabine Hunsiker LT SS 2012

Overview

UNIVERSITÄT DES SAARLANDES

- Introduction: the basic idea
- IBM models: the noisy channel
- Phrase-Based SMT



- Want to learn translation from data
- Data = bitext
- Texts and their translations
- Aligned at sentence level

- Brown et al, "*The Mathematics of Statistical Machine Translation*", Computational Linguistics, 1993
- Tough going

- Fortunately: "*A Statistical MT Workbook*", Kevin Knight, 1999
- These slides are based on Kevin Knight's explanations …

Mary did not slap the green witch

Mary ∅ not slap slap slap the green witch

Mary not slap slap slap NULL the green witch

Maria no daba una bofetada a la verde bruja

Maria no daba una bofetada a la bruja verde

- A generative story
- Given a string in the source language, how can we generate a string in the target language that is a translation
- Components of the story:
  - φ Fertility
  - t Translation (between words)
  - d Distortion (reordering)
  - φ_0 NULL generated words
- Putting them into a model
- Learning the model (parameters) from data

- $P(e)$

- $P(e,f) = P(e) \times P(f)$ if e and f independent

- $P(e,f) = P(e) \times P(f|e)$ if e and f are not independent

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

- $P(e,f) = P(f,e)$

- $P(e|f) \neq P(f|e)$ in general

- $\hat{e} = \arg\max_{e} P(e|f)$

- $P(e|f) = \frac{P(f|e) \times P(e)}{P(f)}$

- $\hat{e} = \arg\max_{e} P(e|f) = \arg\max_{e} \frac{P(f|e) \times P(e)}{p(f)} =$
  $\arg\max_{e} P(f|e) \times P(e)$

- this is the Noisy Channel Model

The Noisy Channel Model

$$\arg\max_{e} P(f|e) \times P(e)$$

- The noisy channel works like this. We imagine that someone has *e* in his head, but by the time it gets on to the printed page it is corrupted by “noise” and becomes *f*. To recover the most likely *e*, we reason about (1) what kinds of things people say any English, and (2) how English gets turned into French. These are sometimes called “source modeling” and “channel modeling.” (Knight, 1999, p.2)

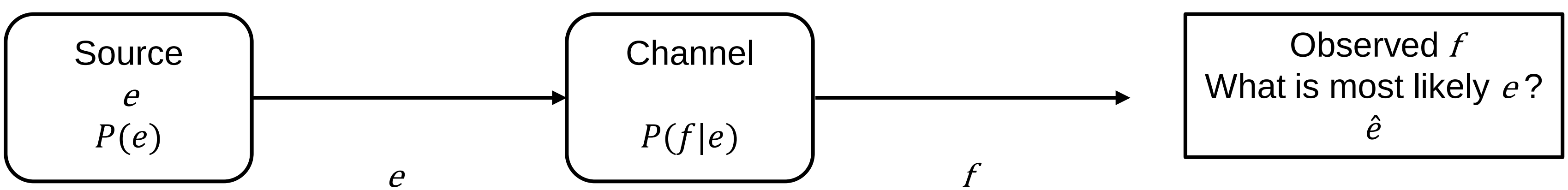
- People use the noisy channel metaphor for a lot of engineering problems, like actual noise on telephone transmissions. (ibid)

The Noisy Channel Model

$$\hat{e} = \arg\max_{e} P(f|e) \times P(e)$$

$P(e)$ the source model, the language model

$P(f|e)$ the channel model, the translation model



Interlude

Chris Dyers slides from MT Marathon 2011 on the Noisy Channel and SMT

Slide: Chris Dyer, MT Marathon 2011

Shannon's theory tells us:

1) the limits of compression
2) why your download is so slow
3) how to recognize speech
4) **how to translate**

Claude Shannon. "A Mathematical Theory of Communication" 1948.

Slide: Chris Dyer, MT Marathon 2011

Claude Shannon. "A Mathematical Theory of Communication" 1948.

$$p(y) \quad p(x|y) \quad p(x)$$

Claude Shannon. "A Mathematical Theory of Communication" 1948.

Slide: Chris Dyer, MT Marathon 2011

$$\boxed{p(y) \quad p(x|y)} \quad p(x)$$

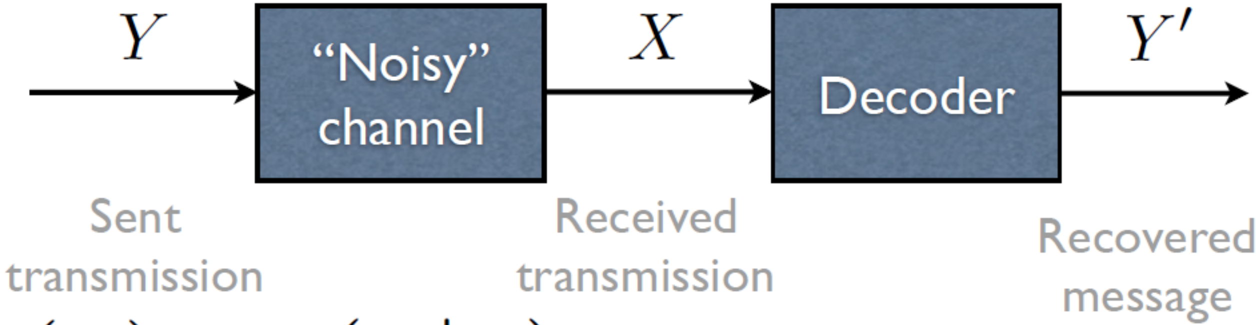
Claude Shannon. "A Mathematical Theory of Communication" 1948.

Slide: Chris Dyer, MT Marathon 2011

$$p(y) \quad p(x|y)$$

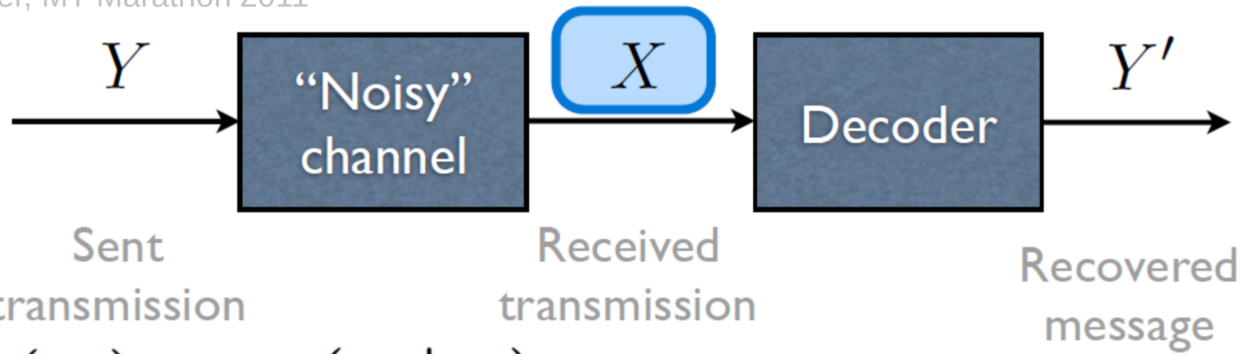
Claude Shannon. "A Mathematical Theory of Communication" 1948.

Slide: Chris Dyer, MT Marathon 2011



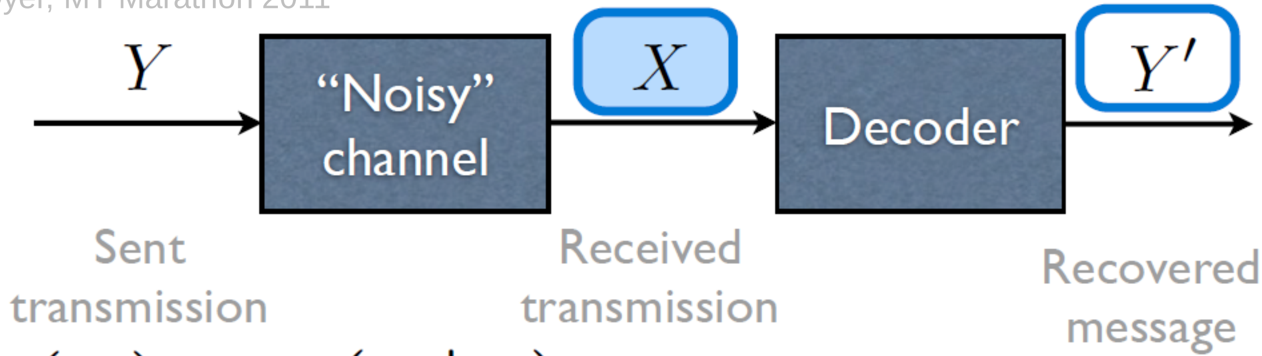
$$p(y) \quad p(x|y)$$

Slide: Chris Dyer, MT Marathon 2011



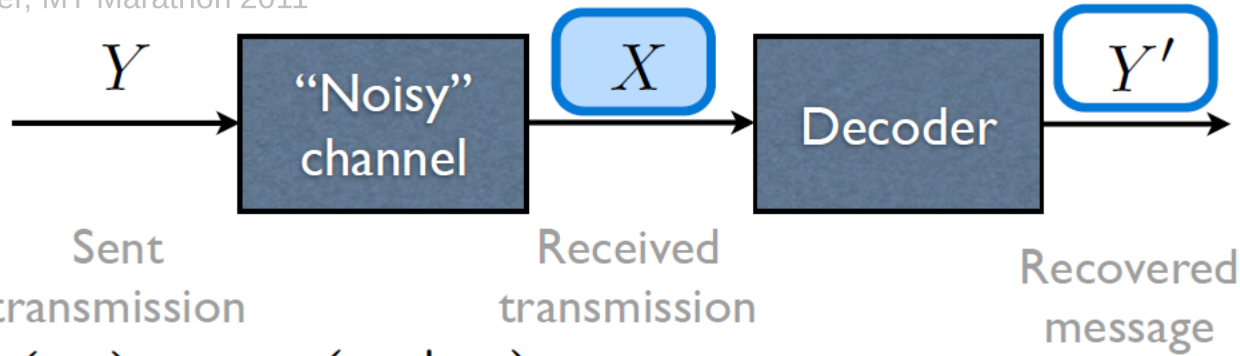
$$p(y) \quad p(x|y)$$

Slide: Chris Dyer, MT Marathon 2011



$$p(y) \quad p(x|y)$$

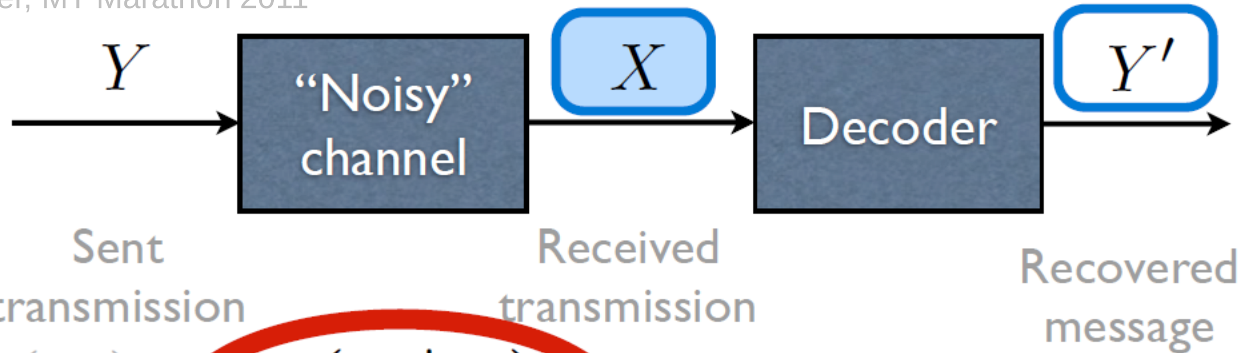
Slide: Chris Dyer, MT Marathon 2011



$$p(y) \quad p(x|y)$$

$$\boxed{y'} = \arg\max_y p(y|x)$$

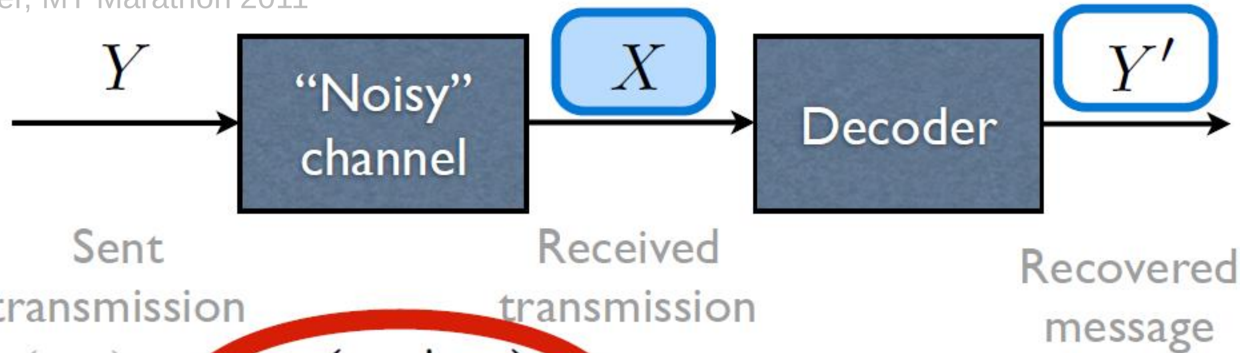
Slide: Chris Dyer, MT Marathon 2011



$$p(y) \quad p(x|y)$$

$$\boxed{y'} = \arg\max_y p(y|x)$$

Slide: Chris Dyer, MT Marathon 2011

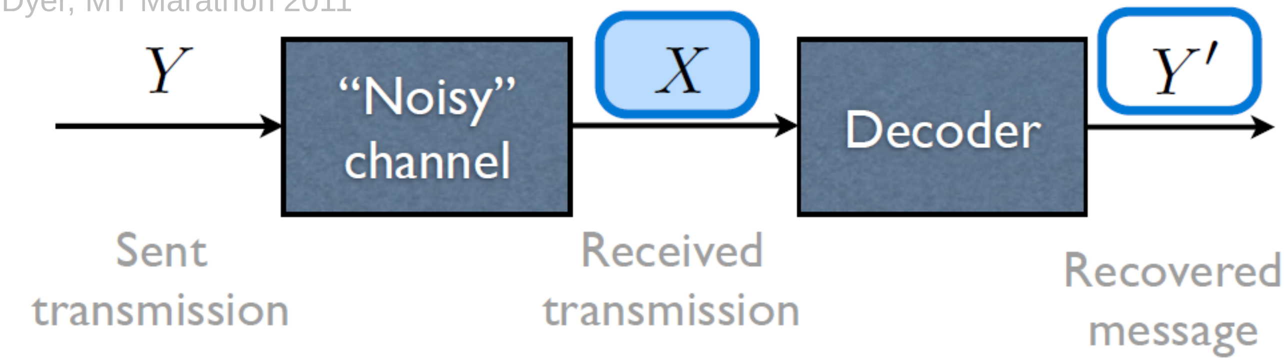


$$p(y) \quad p(x|y)$$

$$\boxed{y'} = \arg\max_{y} p(y|x)$$

I can help.

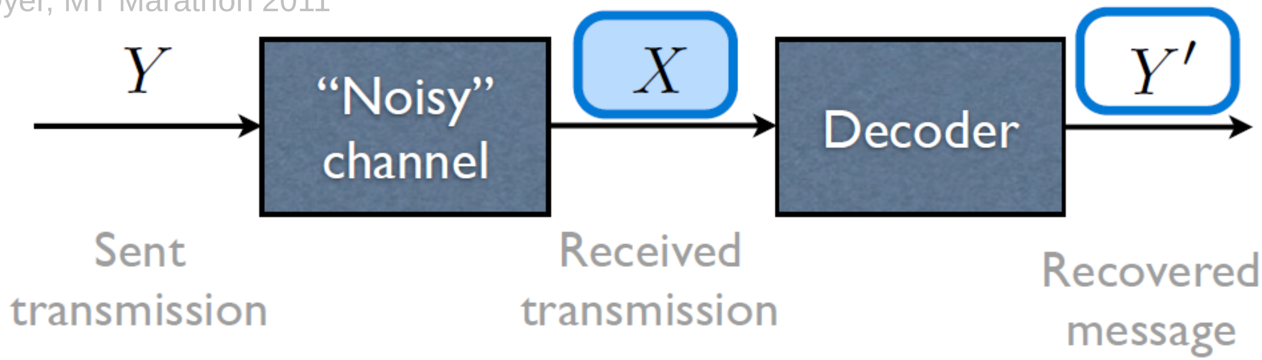
Slide: Chris Dyer, MT Marathon 2011



$$\boxed{y'} = \arg\max_y p(y|x)$$

$$= \arg\max_y \frac{p(x|y)p(y)}{p(x)}$$

Slide: Chris Dyer, MT Marathon 2011

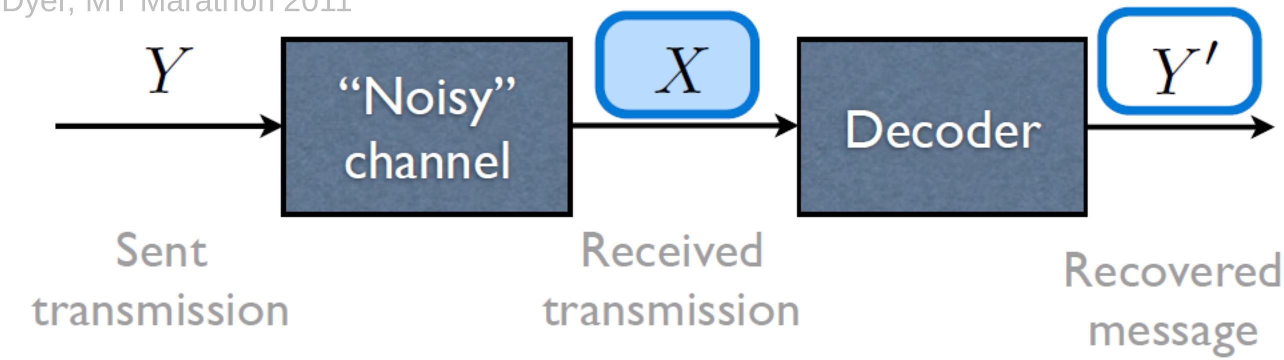


$$\boxed{y'} = \arg\max_y p(y|x)$$

$$= \arg\max_y \frac{p(x|y)p(y)}{p(x)}$$

Denominator doesn't depend on y.

Slide: Chris Dyer, MT Marathon 2011

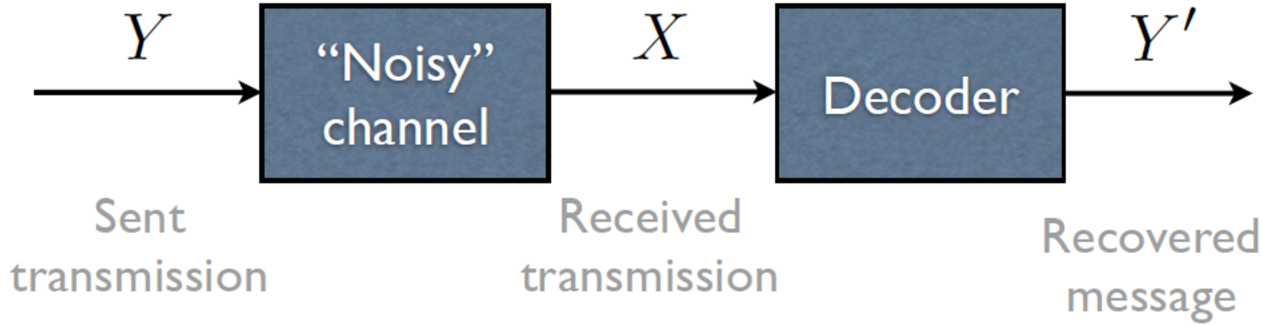


$$\boxed{y'} = \arg\max_y p(y|x)$$

$$= \arg\max_y \frac{p(x|y)p(y)}{p(x)}$$

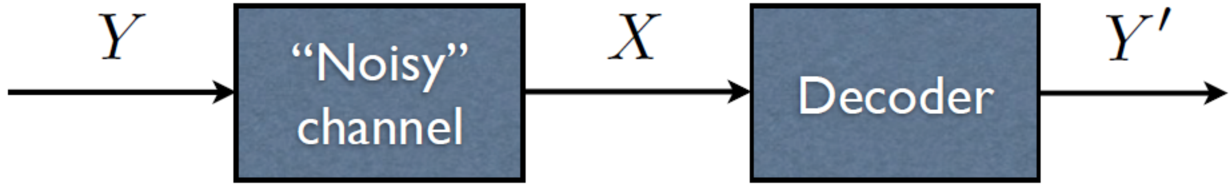
$$= \arg\max_y p(x|y)p(y)$$

Slide: Chris Dyer, MT Marathon 2011



$$y' = \arg\max_y p(x|y)p(y)$$

Slide: Chris Dyer, MT Marathon 2011



Sent transmission

English

Received transmission

"French"

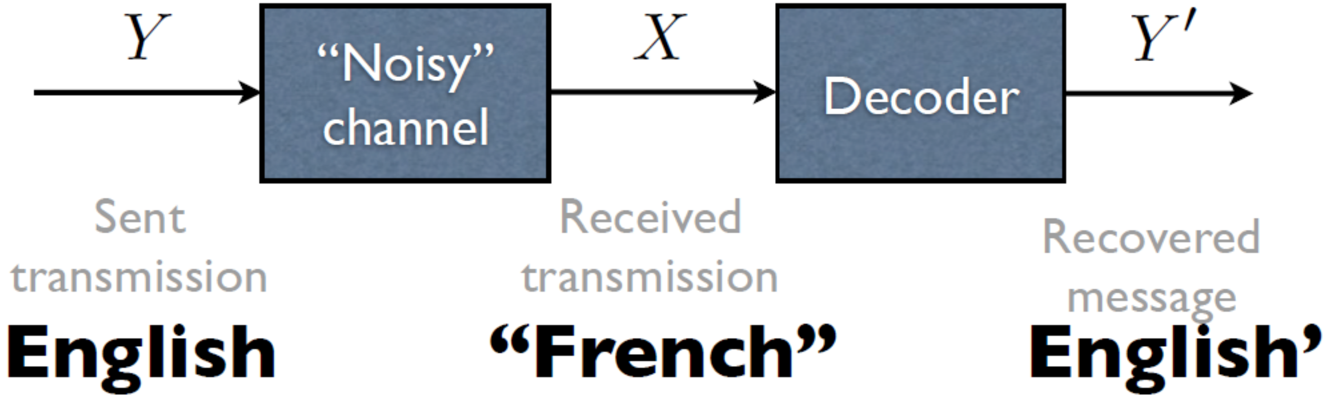
Recovered message

English'

$$\cancel{y' = \arg\max_{y} p(x|y)p(y)}$$

$$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

Slide: Chris Dyer, MT Marathon 2011



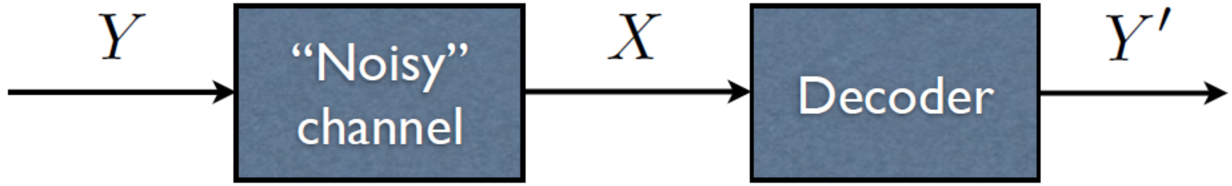
$$\cancel{y' = \arg\max_{y} p(x|y)p(y)}$$

$$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

translation model

language model

Slide: Chris Dyer, MT Marathon 2011



Sent transmission | Received transmission | Recovered message

English | **"French"** | **English'**

$$\cancel{y' = \arg\max_{y} p(x|y)p(y)}$$

$$\mathbf{e}' = \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

translation model | language model

Other noisy channel applications: OCR, speech recognition, spelling correction...

Division of labor

- **Translation model**
  - probability of translation *back* into the source
  - ensures **adequacy** of translation
- **Language model**
  - is a translation hypothesis "good" English?
  - ensures **fluency** of translation

End of Interlude

Back to our slides based on Kevin Knight's 1999 workbook

Translation Modelling

- Remember that translating f to e we reason backwards
- We observe f
- We want to know what e is (most) likely to be uttered and likely to have been translated into f

$$\hat{e} = \arg\max_{e} P(f|e) \times P(e)$$

- Story: replace words in e by French words and scramble them around
- “What kind of a crackpot story is that?” (Kevin Knight, 1999)
- IBM Model 3 ☺

- What happens in translation?
- Actually a lot ….

- EN: `Mary did not slap the green witch`
- ES: `Mary no daba una botefada a la bruja verde`

- But from a purely external point of view

  - Source words get replaced by target words
  - Words in target are moved around (“reordered”)
  - Source and target need not be equally long ….

- So minimally that is what we need to model …

Some parts of the Model

1. For each word e_i in an English sentence $i = (1 \;... l)$, we choose a fertility φ_i. The choice of fertility is dependent solely on the English word in question, nothing else.
2. For each word e_i, we generate φ_i French words: $t(f|e)$. The choice of French word is dependent solely on the English word that generates it. It is not dependent on the English context around the English word. It is not dependent on other French words that have been generated from this or any other English word.
3. All those French words are permuted: $d(\pi_f|\pi_e, l, m)$. Each French word is assigned an absolute target "position slot." For example, one word may be assigned position 3, and another word may be assigned position 2 -- the latter word would then precede the former in the final French sentence. The choice of position for a French word is dependent solely on the absolute position of the English word that generates it.

Translation as String Rewriting

Mary did not slap the green witch

φ

Mary ∅ not slap slap slap the the green witch

t

Maria no daba una bofetada a la verde bruja

d

Maria no daba una bofetada a la bruja verde

Parameters

- We would like to learn the Parameters for fertility, (word) translation and distortion from data
- The parameters look like this
  - $n(3|slap)$
  - $t(maison|house)$
  - $d(5|2,4,6)$
- And they have probabilities associated with them

NULL

UNIVERSITÄT DES SAARLANDES

- One more twist: spurious words
- E.g. function words can appear in target that do not have correspondences in source
- Pretend that every English sentence has NULL word in position 0 and can generate spurious words in target: $t(a|NULL)$
- Longer sentences are more likely to have more spurious words
- $NULL$ therefore doesn't have fertility distribution but a probability p_1 with which it can generate a spurious word after each properly generated word, how many φ_0
- $p_0 = 1 - p_1$ is probability of not tossing in spurious word

NULL Mary did not slap the green witch

Mary ∅ not slap slap slap the green witch

Mary not slap slap slap *NULL* the green witch

Maria no daba una bofetada a la verde bruja

Maria no daba una bofetada a la bruja verde

Model 3

1. For each English word e_i indexed by $i = 1,2,\ldots,l$ choose fertility φ_i with probability $n(\varphi_i | e_i)$.
2. Choose the number φ_0 of "spurious" French words to be generated from $e_0 = NULL$, using probability p_1 and the sum of fertilities from step 1.
3. Let m be the sum of fertilities for all words, including $NULL$.
4. For each $i = 1,2,\ldots,l$ and each $k = 1,2,\ldots,\varphi_i$ choose a French word $\tau_{i,k}$ with probability $t(\tau_{i,k} | e_i)$.
5. For each each $i = 1,2,\ldots,l$ and each $k = 1,2,\ldots,\varphi_i$ choose target French position $\pi_{i,k}$ with probability $d(\pi_{i,k} |\ i,l,m)$.
6. For each $k = 1,2,\ldots,\varphi_i$ choose a position $\pi_{0,k}$ from the φ_0 - $k + 1$ remaining vacant positions in $1,2,\ldots,m$ for a total probability of $1/\varphi_0$!.
7. Output the French sentence with words $\tau_{i,k}$ in positions $\pi_{i,k}$ $(0 \le i \le l, 1 \le k \le \varphi_i)$.

Another Interlude

Some slides from Sabine Hunsieker

Sources for Information

UNIVERSITÄT DES SAARLANDES

- MT in general, history:
  - http://www.MT-Archive.info: Electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools, regularly updated, contains over 3300 items
  - Hutchins, Somers: An introduction to machine translation. Academic Press, 1992, available under http://www.hutchinsweb.me.uk/IntroMT-TOC.htm
- MT systems:

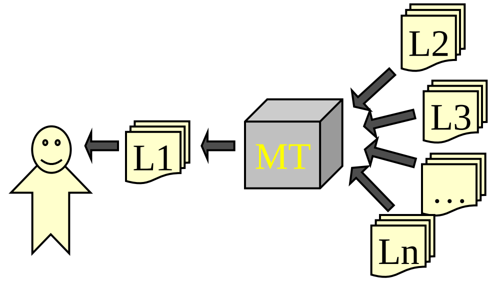
  Compendium of Translation Software, see
  http://www.hutchinsweb.me.uk/Compendium.htm
- Statistical Machine Translation:

  See www.statmt.org

  Book by Philipp Koehn is available in the coli-bib

Use cases and requirements for MT

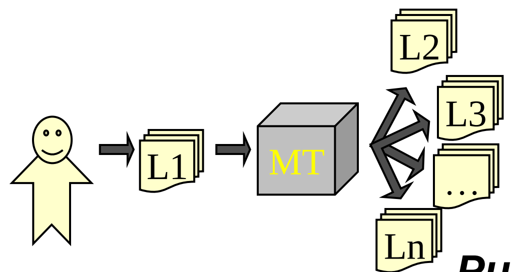
a) MT for assimilation
„inbound“



Robustness
Coverage

Daily throughput of online-MT-Systems > 500 M Words

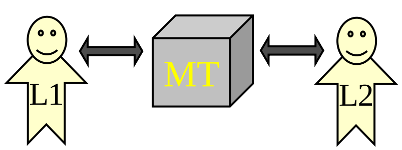
b) MT for dissemination
„outbound“



Textual quality

Publishable quality can only be authored by humans; Translation Memories & CAT-Tools mandatory for professional translators

c) MT for direct communication



Speech recognition, context dependence

Topic of many running and completed research projects (VerbMobil, TC Star, TransTac, …)
US-Military uses systems for spoken MT

On the Risks of Outbound MT

Some recent examples

'I am not in the office at the moment. Please send any work to be translated'

Motivation for rule-based MT

UNIVERSITÄT DES SAARLANDES

- Good translation requires knowledge of linguistic rules
  - …for understanding the source text
  - …for generating well-formed target text

- Rule-based accounts for certain linguistic levels exist and should be used, especially for
  - Morphology
  - Syntax

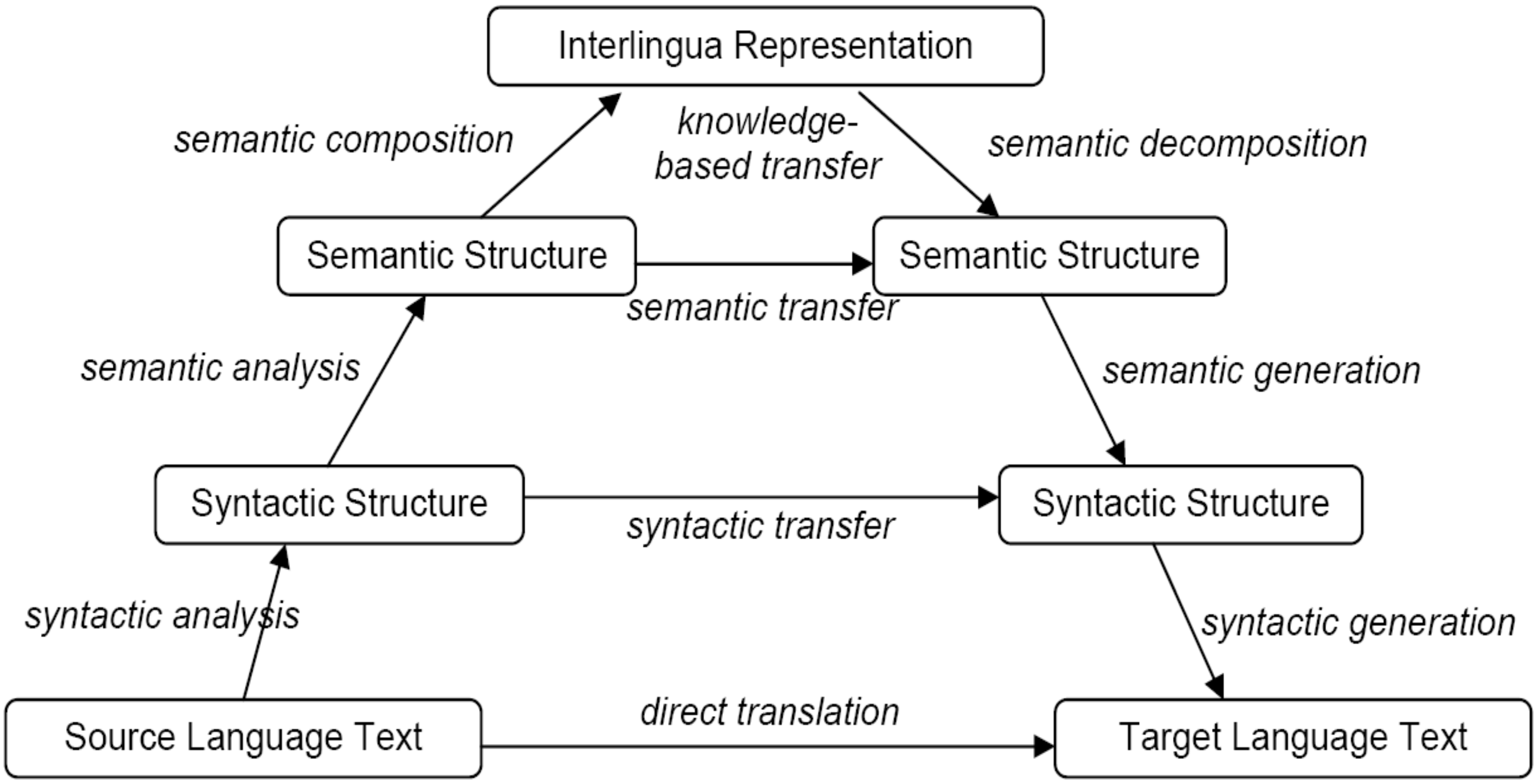
- Writing one rule is better than finding hundreds of examples, as the rule will apply for new, unseen cases

- Following a set of rules can be more efficient than search for the most probable translation in a large statistical model

Possible (rule-based) MT architectures

UNIVERSITÄT DES SAARLANDES

The „Vauquois Triangle“



Motivation for statistical MT

- Good translation requires knowledge and decisions on many levels
  - syntactic disambiguation (POS, attachments)
  - semantic disambiguation (collocations, scope, word sense)
  - reference resolution
  - lexical choice in target language
  - application-specific terminology, register, connotations, good style …
- Rule-based models of all these levels are very expensive to build, maintain, and adapt to new domains
- Statistical approaches have been quite successful in many areas of NLP, once data has been annotated
- Learning from existing translation will focus on distinctions that matter (not on the linguist's favorite subject)
- Translation corpora are available in rapidly growing amounts
- SMT *can* integrate rule-based modules (morphologies, lexicons)
- SMT *can* use feed-back for on-line adaptation to domain and user preferences

History of SMT and Important Players I

- 1949: Warren Weaver: *the translation problem can be largely solved by "statistical semantic studies"*
- 1950s..1970s: Predominance of rule-based approaches
- 1966: ALPAC report: general discouragement for MT (in the US)
- 1980s: example-based MT proposed in Japan (Nagao), statistical approaches to speech recognition (Jelinek e.a. at IBM)
- Late 80s: Statistical POS taggers, SMT models at IBM, work on translation alignment at Xerox (M. Kay)
- Early 90s: many statistical approaches to NLP in general, IBM's Candide claimed to be as good as Systran
- Late 90s: Statistical MT successful as a fallback approach within Verbmobil System (Ney, Och). Wide distribution of translation memory technology (Trados) indicates big commercial potential of SMT
- 1999 Johns Hopkins workshop: open source re-implementation of IBM's SMT methods (GIZA)

History of SMT and Important Players II

- Since 2001: DARPA/NIST evaluation campaign (XYZ → English), uses BLEU score for automatic evaluation
- Various companies start marketing/exploring SMT:
  language weaver, aixplain GmbH, Linear B Ltd., esteam, Google Labs
- 2002: Philipp Koehn (ISI) makes EuroParl corpus available
- 2003: Koehn, Och & Marcu propose *Statistical Phrase-Based MT*
- 2004: ISI publishes Philipp Koehn's SMT decoder *Pharaoh*
- 2005: First SMT workshop with shared task
- 2006: Johns Hopkins workshop on OS factored SMT decoder Moses, Start of EuroMatrix project for MT between all EU languages, Acquis Communautaire (EU laws in 20+ languages) made available
- 2007: Google abandons Systran and switches to own SMT technology
- 2009: Start of EuroMatrixPlus *"bringing MT to the user"*
- 2010: Start of many additional MT-related EU projects (Let's MT, ACCURAT, …)