Statistical Methods for MT: IBM models

Prasanth Kolachina

Statistical methods for NLP

March 10^{th} 2016

Outline

What is *M.* Translation?

Machine Translation

- Translation is task of transforming text in one language to another language
  - interpretation of meaning
  - preservation of meaning and structure in original text
- Importance of context in interpretation and translation

  There is nothing outside the text.

  – *Jacques Derrida, "Of Grammatology" (1967)*

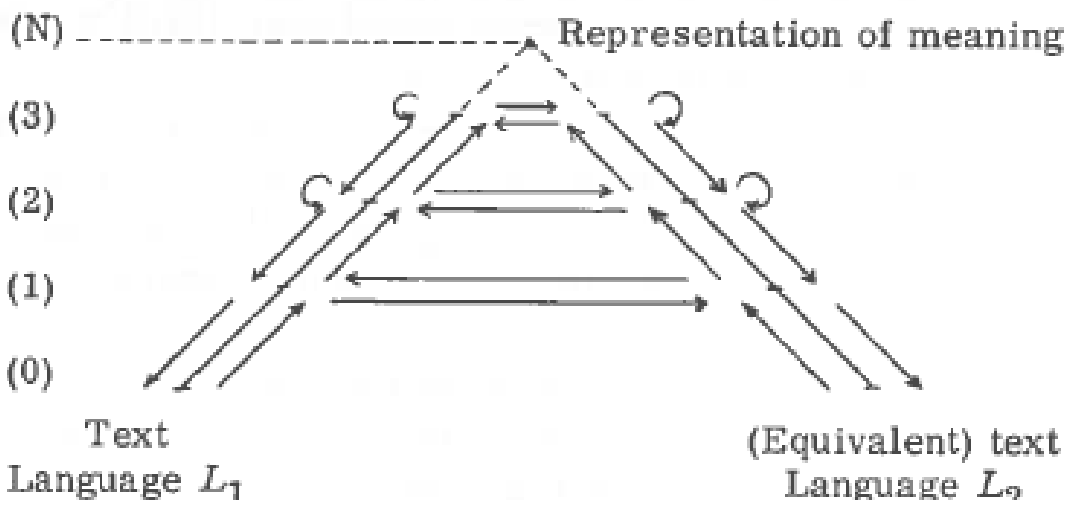
- Can this transformation process be automatized?
  - Machine Translation
  - To what extent is it possible?

Origins of Mechanical Translation

- First ideas from information theory
  - "Translation memorandum" (Weaver [1955])
  - Essentially, decode the information in one language and re-encode the same in target language
- Early attempts to translate using a bilingual dictionary
- Information encoding in text is more complex than simple word meanings
  - Encoded at different levels of "linguistic analysis"
  - Morphology, Syntax, Semantics, Discourse and Pragmatics
- ALPAC report led to the creation of Computational Linguistics
  - Advanced research in both Linguistics and Computer Science
  - E.g. Quick sort algorithm
- Was originally called Mechanical Translation!

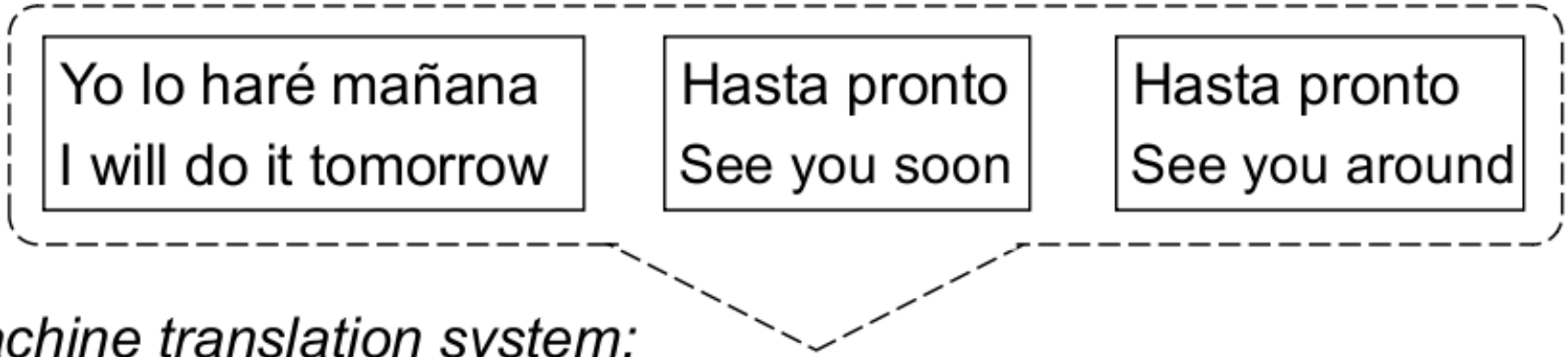
Formalizing approaches to MT

- Post ALPAC report of 1966
  - Role of formal grammars and algorithms for MT (Vauquois [1968])
  - Natural Language Understanding, Natural Language Generation

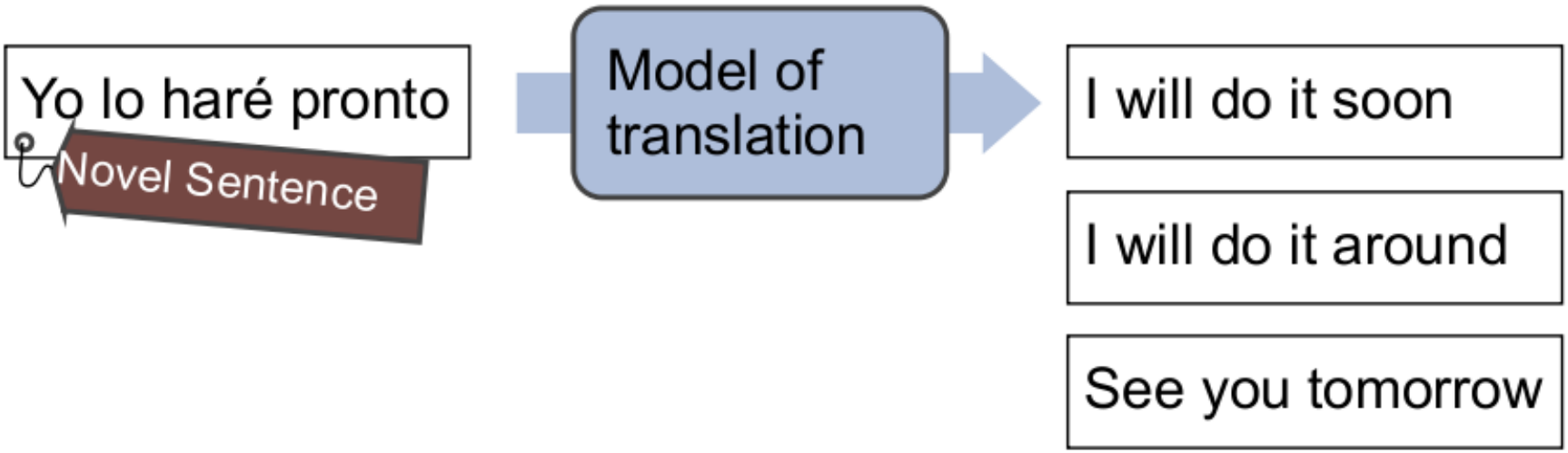


20 years later- Corpus-based MT

Sentence-aligned parallel corpus:



Machine translation system:



[1] Example from Petrov [2012]

Statistical MT

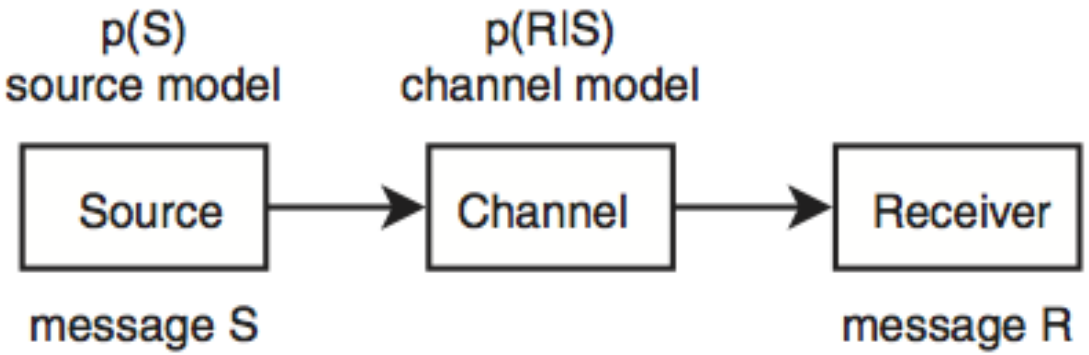
Noisy-Channel model

- Warren Weaver's "memorandum"

  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".

  – Weaver [1955]

- Translation in English (S) can be "reconstructed" for a sentence in Russian (R) using
  - a source model, i.e. language model and
  - a channel model, i.e. translation model



Statistical MT

- Translation from a foreign (F) language to English (E) is a search problem
- Find the most likely English translation using the statistical model

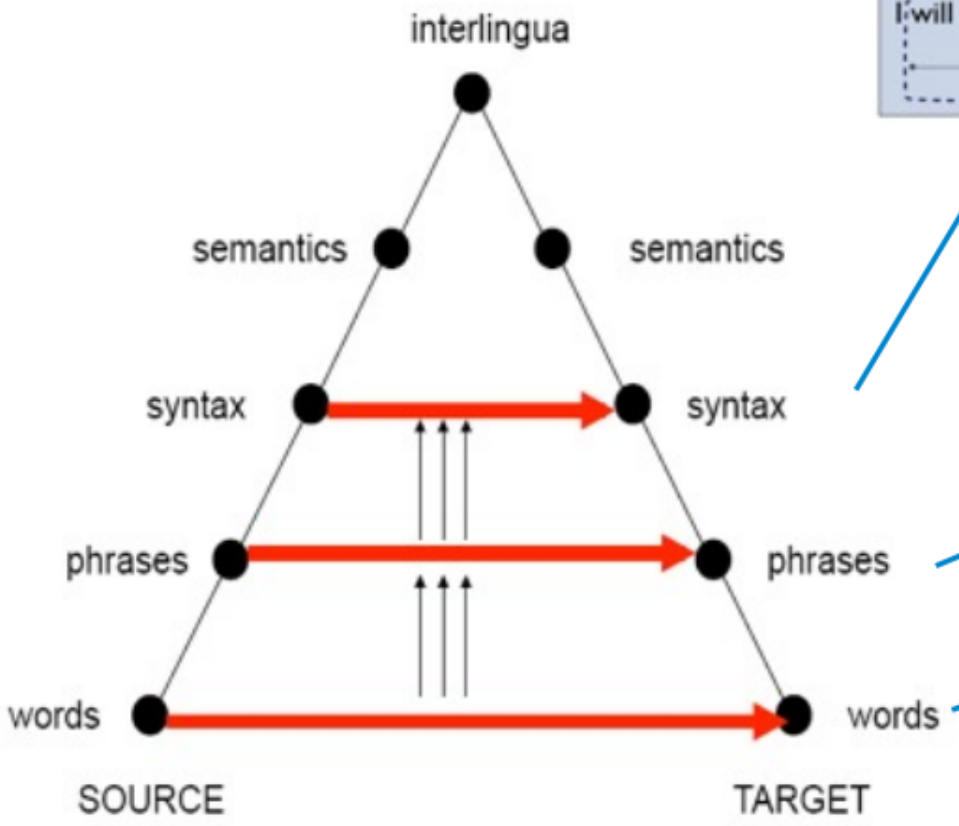
$$\hat{E} = \arg\max_{E} \quad \mathbf{P}(\mathbf{E}|\mathbf{F})$$

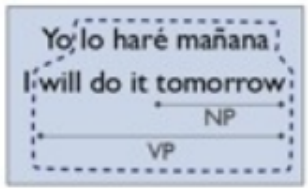
- Bayes rule

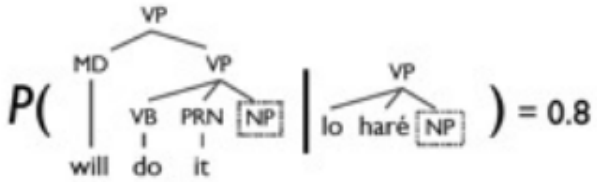
$$\boxed{\hat{E} = \arg\max_{E} \mathbf{P_{TM}}(\mathbf{F}|\mathbf{E}) * \mathbf{P_{LM}}(\mathbf{E})}$$

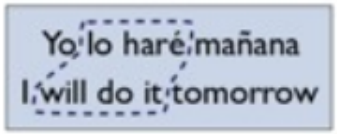
- Two primary components in the model
  - Translation model $\mathbf{P_{TM}} \approx$ channel model
  - Language model $\mathbf{P_{LM}} \approx$ source model

Formalizing approaches to SMT

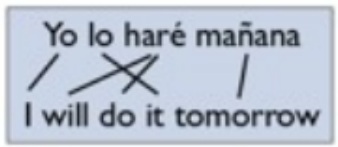








English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2



English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

[2] Example from Petrov [2012]

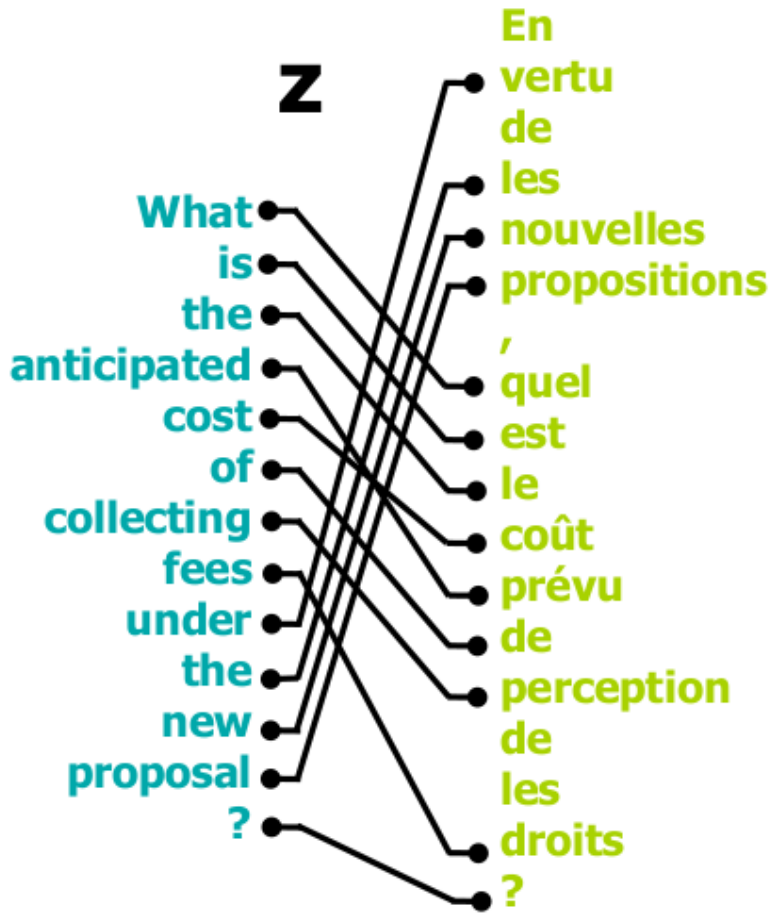
Word-level MT

x

What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?

z



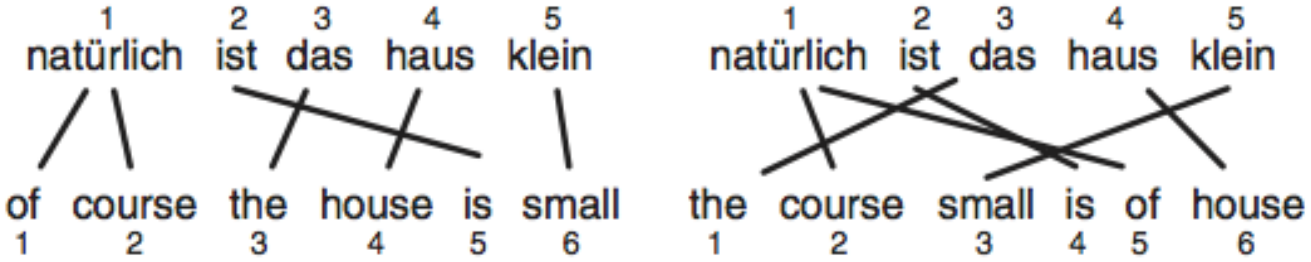
[3] Example from Petrov [2012]

IBM models

- Different models to capture regular variations across language
  - morphology
  - word order
- Models 1-4 for $\mathbf{P_{TM}}$
- How to
  - estimate parameters using Expectation Maximization
  - translate new sentences using these models i.e. decoding
- Model 1 today.

Nuts and *Bolts* of the IBM Models

IBM Model 1

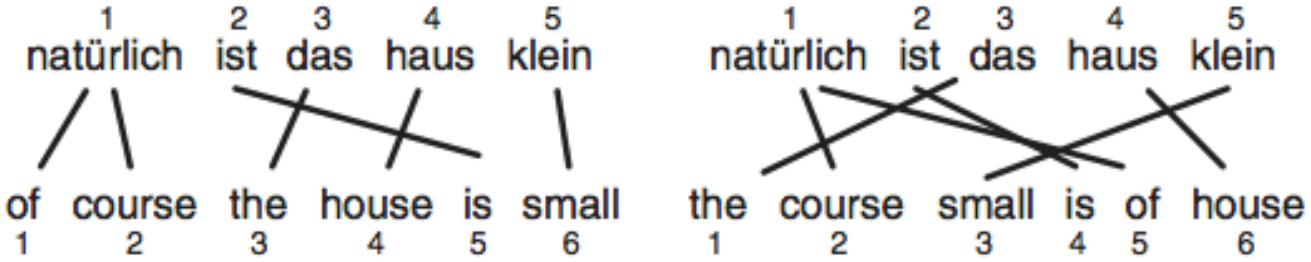


- Given sentence pairs annotated with word-alignments
  - Estimate translation probability distributions $p(f|e)$
  - MLE based on counts
  - This is called the *lexical translation table*

$$t(\text{haus/house}) = \frac{C(\text{haus, house})}{C(\text{house})}$$

$$t(\text{das/the}) = \frac{C(\text{das, the})}{C(\text{the})}$$

IBM Model 1

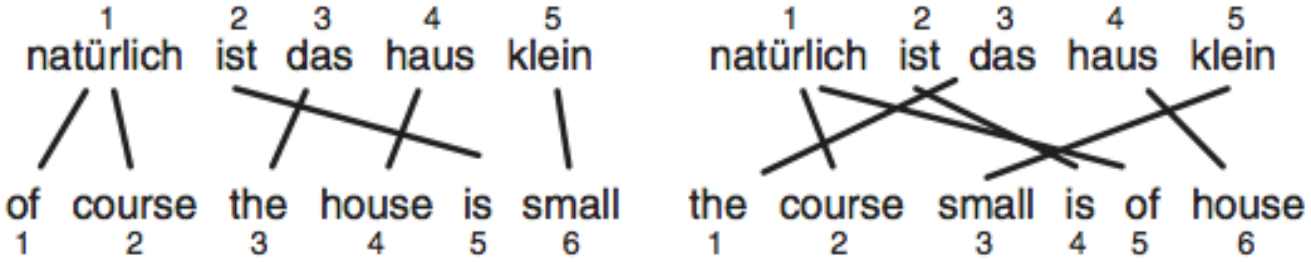


- Given sentence pairs with a lexical translation table
  - there are K alignment sequences that generates the translation pair
  - probability of a alignment sequence a

$$P_{align}(a|e,f) = \frac{P(a,e|f)}{\sum_1^K P(a,e|f)}$$

$$P(a,e|f) = \prod_{i=1}^{m} t(e_i|f_{a(i)})$$

IBM Model 1



- Given sentence pairs with P_{align}
  - Estimate translation probability distributions $p(f|e)$
  - MLE based on "soft" counts P_{align}

$$t(\text{haus}/\text{house}) = \frac{C'(\text{haus, house})}{C'(\text{house})}$$

$$t(\text{das}/\text{the}) = \frac{C'(\text{das, the})}{C'(\text{the})}$$

Parameter estimation in IBM Models

- lexical translation $\rightarrow$ alignments $\rightarrow$ lexical translation
  - alignments are unobserved i.e. latent variables
- Recall the EM algorithm from last lecture
  - (P. Dempster et al. [1977])
  - EM estimates the distributions when hidden variables are present
- E-step estimates P_{align} (expectation)
- M-step estimates the lexical translation table $t(f|e)$ (maximization)

EM in Model 1

- Expectation step:
  - Given probability distribution $t(f|e)$ from previous iteration
  - Estimate the probability of different alignment sequences

$$P_{align}(a|e,f) = \frac{P(a,e|f)}{P(e|f)} = \frac{P(e,a|f)}{\sum_{a'} P(e,a'|f)}$$

  - $P(a,e|f)$ can be computed from $t(f|e)$.

$$P(a,e|f) = \prod_{j=1}^{n} t(e_j|f_{a(j)})$$

$$P_{align}(a|f,e) = \frac{\prod_{j=1}^{n} t(e_j|f_{a(j)}}{\sum_{i=0}^{m} \prod_{j=1}^{n} t(e_j|f_{a(j)})}$$

EM in Model 1

- Maximization step
  - Using P_{align} for each of K possible alignment sequences
  - Compute *new* lexical translation table
  - Use P_{align} as "soft" counts
    - Each instance of an alignment is counted as the probability associated with that sequence

$$c(e, f) = \sum_{a} P_{align}(a|f, e)$$

Practical issues

- How to implement this EM for IBM models?
  - Initialize parameter tables at random (or *uniform*?)
  - Estimate the probability of hidden alignments P_{align}
  - Estimate new parameter table values t using P_{align}
  - Iterate over these two steps until EM reaches convergence
    - converge when entropy of the model does not change
- What is K? The number of possible alignments
  - $(n+1)^m$
- EM will converge for model 2 Collins [2012]
- The result can be local optimum rather than "real" solution

IBM Models

- Modeling $\mathbf{P_{TM}}$
- Different parameter are defined to explain translation process
  - lexical translation $t(f|e)$ –model 1
  - distortion q –model 2
  - fertility n –model 3
  - relative distortion q' –model 4
- $t(f|e)$ for the current discussion

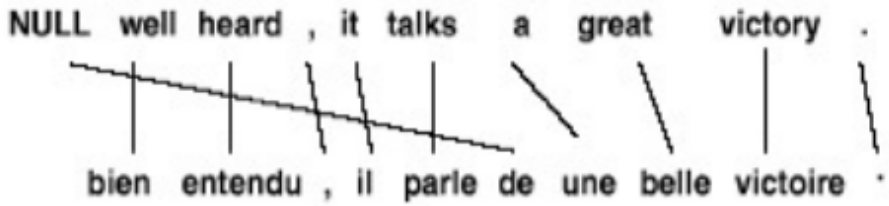
Decoder

- Given a translation model $\mathbf{P_{TM}}$ and a language model for target language $\mathbf{P_{LM}}$
  - find the most "likely" translation for a source sentence
- An intractable problem: no exact solution
  - Maximize over all possible translations
  - Each translation can be generated by many underlying alignments
  - Sum over all such plausible alignments
- Number of plausible permutations and alignments are exponential in sentence length
- Inexact search instead of exact search
  - approximations make decoding tractable
  - greedy decoding
  - beam-search

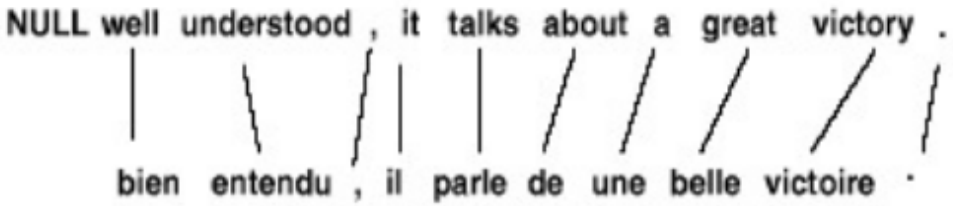
Greedy decoder

- Start by assigning each word its most probable translation
  - hypothesis
- Compute the probability of the hypothesis
  - scores from both $\mathbf{P_{TM}}$ and $\mathbf{P_{LM}}$
- Make mutations to the hypotheses until no difference in probability scores (Turitzin [2005])
- What are plausible mutations
  - Change translation options for each word
  - Add new words to hypothesis or remove existing words
  - Moving words around inside the hypothesis
    - swap non-overlapping segments

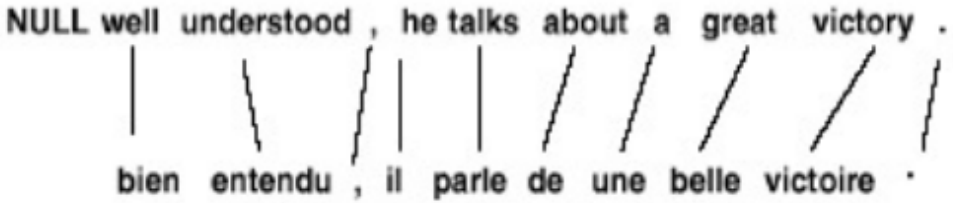
Decoding example



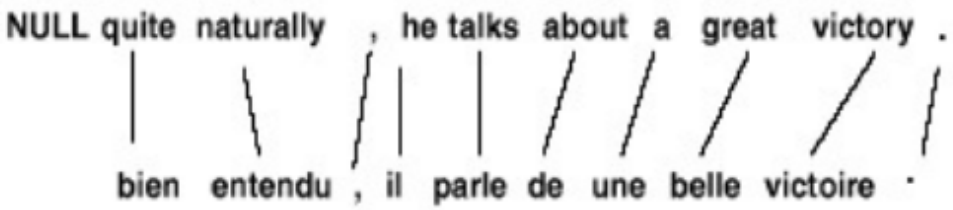
translateTwoWords(2,understood,0,about)



translateOneWord(4,he)



translateTwoWords(1,quite,2,naturally)



Beyond Word models in SMT

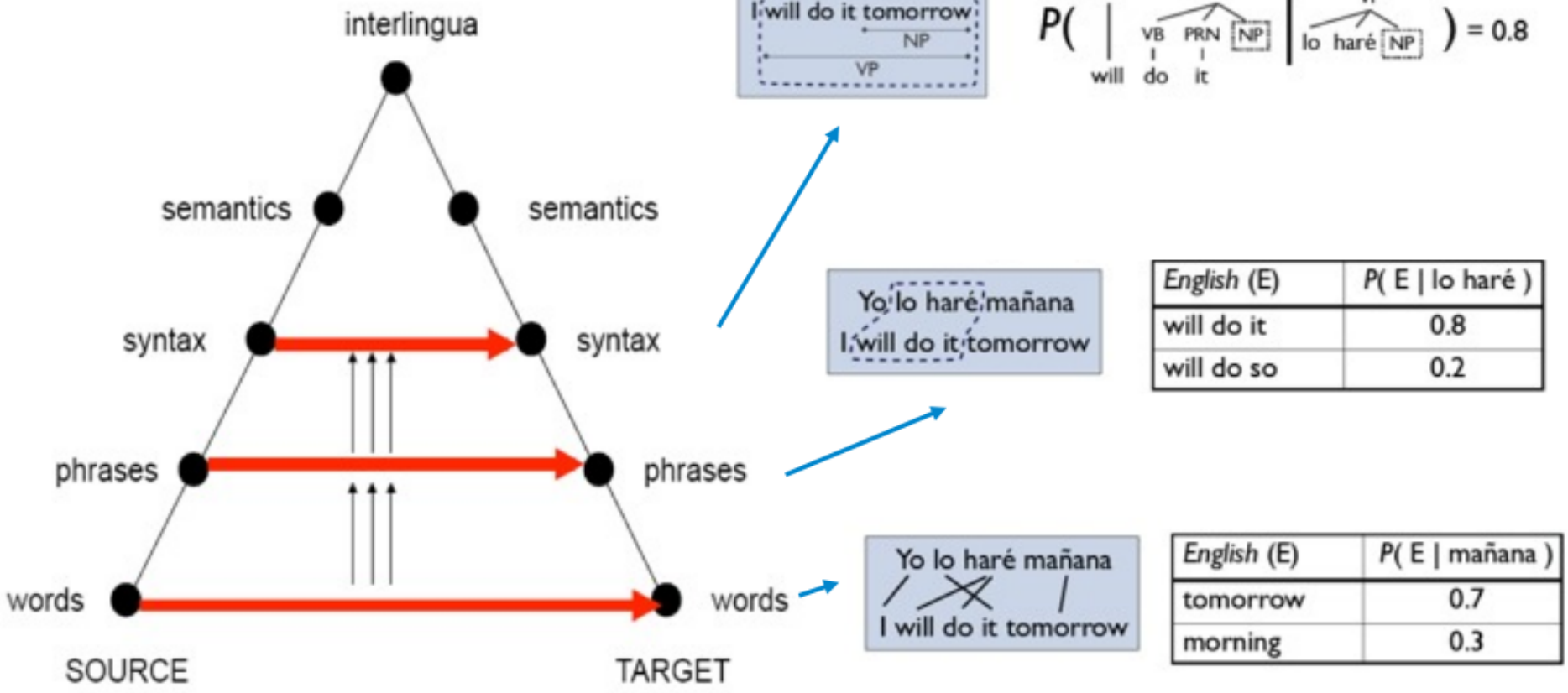
Shortcomings of IBM Models

- Simplifying assumptions in model formulation (Brown et al. [1993])
- Lack of context in predicting likely translation of a word
  1. The **ball** went past the **bat** and hit the stumps in the last **ball** of the innings.
  2. The **bat** flew out of the cave with wings as black as night itself.
  3. They danced to the music all night at the **ball**.
- Not very different from dictionary lookup to translate
- Discarding linguistic information encoded in a sentence
  - Morphological variants
  - Syntactic structure like part-of-speech tags
- Multi-word concepts
  - **break the ice**
  - **liven up**

Extending words to phrases

- Phrasal translations rather than word translations (Koehn et al. [2003])
- Simple way to incorporate local context into translation model
- Phrase pairs are extracted using alignment template
  - Word alignments are used to extract "good" phrase pairs
  - Reordering at phrase level instead of word reorderings
- Notion of phrase is not defined linguistically
  - any n-gram in the language is a phrase
- State-of-art models

Reiterating ..



English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2

English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

[4] Example from Petrov [2012]

Encoding Linguistic Information

- Various levels of linguistic information
- Morphology: gender, number or tense
  - Factored phrase models
- Syntax: syntactic reordering between language pairs
  - regular patterns for a language pair
  - for e.g. adjectives in English and French or
  - Clause reordering between English and German
  - Syntax-based SMT
- Other information
  - Semantics, Discourse, Pragmatics
- All of these are open research problems !!

Evaluating MT

- Evaluation criteria
  - fluency of translations
  - adequacy i.e. translations preserving meaning
- Human judgements are most reliable
  - Nießen et al. [2000]
  - Very expensive and time-consuming
  - Variation in judgements
- Automatic evaluation metrics
  - Compute similarity of translations to reference translations
  - BLEU, NIST (A. Papineni et al. [2002]) and many more
  - Choice of metric varies depending on application requirements
- How to interpret evaluation scores?

Next?

- VG assignment (optional)
  - Implement IBM Model 1
- Help session next week
- Interested further in MT
  - Feel free to contact Richard or me :-)

Reading List

- Lecture notes on IBM Models 1 and 2 (Collins [2012])
- Text book on SMT: Chapter 4 (Koehn [2010])
- Tutorial by Kevin Knight (Knight [1999])
  - Gives a detailed explanation of the math behind IBM Models

References I

A. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P02-1040`.

Brown, Peter E., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19:263–311. URL `http://aclweb.org/anthology-new/J/J93/J93-2003`.

Collins, Michael. 2012. Statistical Machine Translation: IBM Models 1 and 2.

Knight, Kevin. 1999. A Statistical MT Tutorial Workbook. URL `http://www.isi.edu/natural-language/mt/wkbk-rw.pdf`.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of Human Language Technologies: The 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 48–54. Edmonton, Canada: Association for Computational Linguistics. URL `http://aclweb.org/anthology-new/N/N03/N03-1017`.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Herrmann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the Second Conference on International Language Resources and Evaluation (LREC'00)*, 39–45. Athens, Greece: European Language Resources Association (ELRA).

P. Dempster, Arthur, Laird M. Nan, and Bruce Rubin Donald. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39:1–38. URL `http://www.jstor.org/stable/2984875`.

Petrov, Slav. 2012. Statistical NLP.

Turitzin, Michael. 2005. SMT of French and German into English Using IBM Model 2 Greedy Decoding.

Vauquois, Bernard. 1968. A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Mechanical Translation. In *Proceedings of IFIP Congress*, 1114–1122. Edinburgh.

Weaver, Warren. 1955. Translation. Technical report, Cambridge, Massachusetts.