Natural Language Processing
SoSe 2014

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Machine Translation

Dr. Mariana Neves

July 2nd, 2014

(based on the book of Jurafski and Martin 2009)

Outline

- Task
- Classical MT
- Statistical MT
- Alignment in MT
- Evaluation

Outline

- Task
- Classical MT
- Statistical MT
- Alignment in MT
- Evaluation

Machine Translation

- Use of computer to automate translation from one language to another

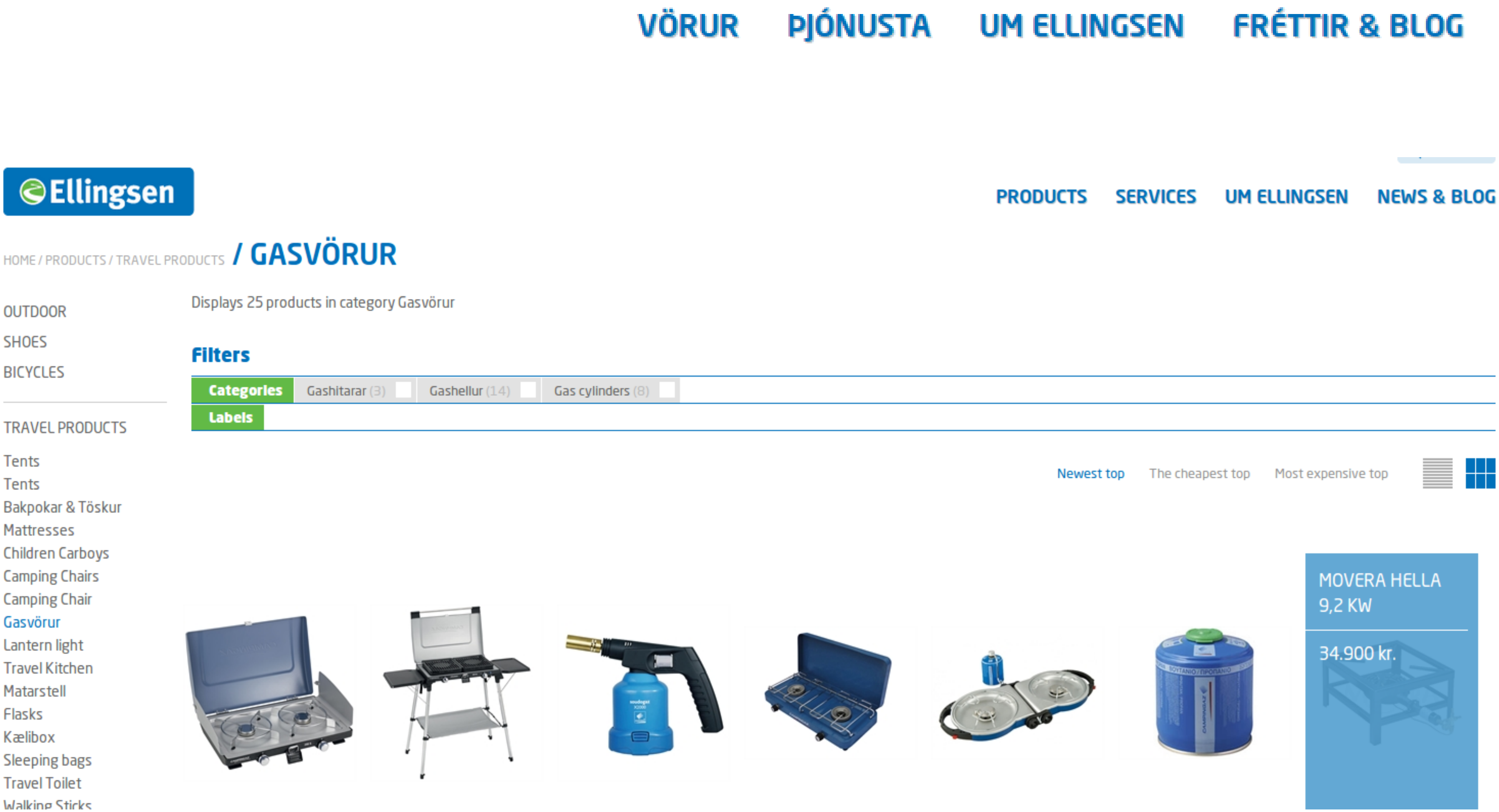
Machine Translation

- C1: DAIYU ALONE ON BED TOP THINK BAOCHAI
- E1: As she lay there alone Daiyu's thoughts turned to Baochai .
- C3: CLEAR COLD PENETRATE CURTAIN
- E3: The coldness penetrated the curtains of her bed .
- C4: NOT FEELING FALL DOWN TEARS COME
- E4: Almost without noticing it she had began to cry .

Machine Translation

- Applications of current MT computational models
  - Rough translation
  - Human post-editor
  - Fully automatic, high-quality translation (FAHQT)
    - small sublanguage domains

Rough translation



Human post-editing

- Computer-aided human translation (CAHT or CAT)
  - Software manual
  - PhD thesis

Sublanguage domain (FAHQT)

- Domains with limited vocabulary
  - Weather forecast
    - Cloudy with a chance of showers today
    - Outlook for Friday: Sunny
  - Air travel queries
  - Appointment scheduling
  - Restaurant recommendation
  - Recipes

Challenges of MT

- Typology: study of cross-linguistic similarities and differences
  - Morphology
    - Agglutinative: Turkish
    - Fusion: Russian
      - „solom“: (table-SG-INSTR-DECL1)
        - Instrumental, singular, first declension

Challenges of MT

- Typology: study of cross-linguistic similarities and differences
  - Syntax: order of verbs, subjects and objects
    - SVO: German, French, English, Mandarin
    - SOV: Hindi, Japanese
    - VSO: Irish, Arabic, Biblical Hebrew

    - English: „He adores listening to music."
    - Japanese: „he music to listening adores"
    - German: „he adores music to listen"

Challenges of MT

- Typology: study of cross-linguistic similarities and differences
  - Argument structure and linking
    - Head-marking
    - Dependent-marking

    - English: „the man's house"
    - Hungarian: „the man house-his"

Challenges of MT

- Typology: study of cross-linguistic similarities and differences
  - Verbs and satellite particles (direction, motion, etc.)
  - Verb-framed
    - Spanish: „The bottle exited floating.“ (verb)
    - approach, exit, reach, enter
    - Japanese, Tamil, Romance, Semitic, Mayan
  - Satellite-framed
    - English: „The bottle floated out.“ (particle)
    - crawl out, float off, jump down, run after
    - English, Swedish, Russia, Indi, Farsi

Challenges of MT

- Typology: study of cross-linguistic similarities and differences
  - Pronouns omission
    - Pro-drop:
      - English: [I] am reading a book.
      - Spanish: Estoy leyendo un libro.
    - Referential density
      - Cold: more inferential work to recover antecedents
        - Japanese, Chinese
      - Hot: more explicit and easier
        - Spanish

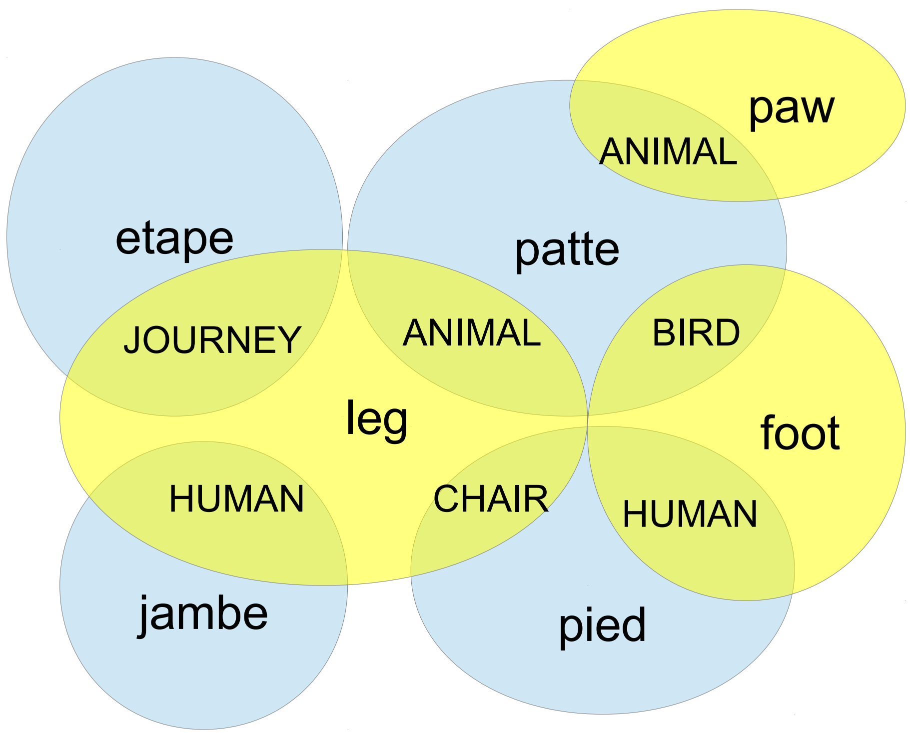
Challenges of MT

- Other divergences:
  - Position of adjectives
    - English: „green witch“
    - Spanish: „witch green“
  - Chinese relative clauses (in relation to English)
  - Cultural aspects, e.g., calendars and dates
    - British English: DD/MM/YY
    - American English: MM/DD/YY
    - Japanese: YYMMDD

Challenges of MT

- Lexical
  - Word sense disambiguation
  - Homonymy
    - wall (Wand), wall (Mauer)
  - Polysemy
    - to know (knowing a fact) : wissen
    - to know (familiarity with a person/location): kennen
  - Grammar
    - English: „She likes to sing“
    - German: „Sie singt gern.“
  - Lexical gap
    - No „privacy“ word in Japanese

Challenges of MT



(figure derived from Jurafsky and Martin 2009)

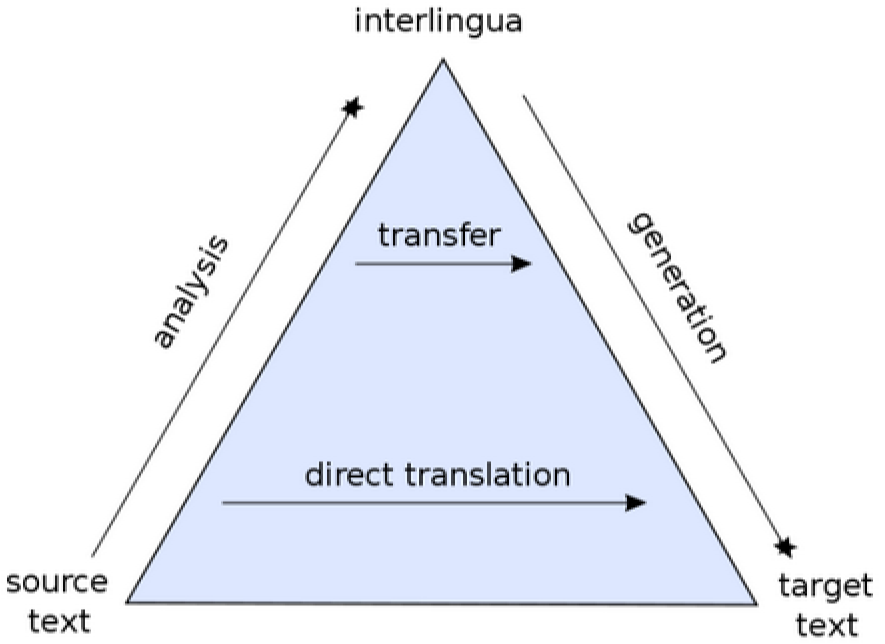
Outline

- Task
- Classical MT
- Statistical MT
- Alignment in MT
- Evaluation

Classical MT

- Direct
  - Word-by-word translation
- Transfer
  - Source language parse structure to target language parse structure
- Interlingual
  - Analysis using an abstract meaning representation

Vauquois triangle



(http://kuiwon.wordpress.com/2013/05/24/an-overview-of-automated-machine-translation-the-ruled-based-approach/)

Direct translation

- Word-by-word translation

- No intermediate structures, except for shallow morphological analysis
  - e.g., verb tenses, negations
- Use of bilingual dictionaries
- Followed by reordering rules
  - e.g., moving adjectives after nouns

Direct translation

- English to Spanish:
  - English: „Mary didn't slap the green witch."
  - Spanish: „Maria no dió una bofetada en la bruja verde"

    Mary not gave a slap to the witch green

  - Input: Mary didn't slap the green witch.
  - After Morphology: Mary DO-PAST not slap the green witch
  - After lexical transfer: Maria PAST no dar una bofetada a la verde bruja
  - After local reordering: Maria no dar PAST una bofetada a la bruja verde
  - After Morphology: Mariana no dio una bofetada a la bruja verde

Direct translation

- Shortcomings:
  - Need complex bilingual dictionaries
    - „slap“ → „dar una bofetada a“
  - Not realible for long distance reordering

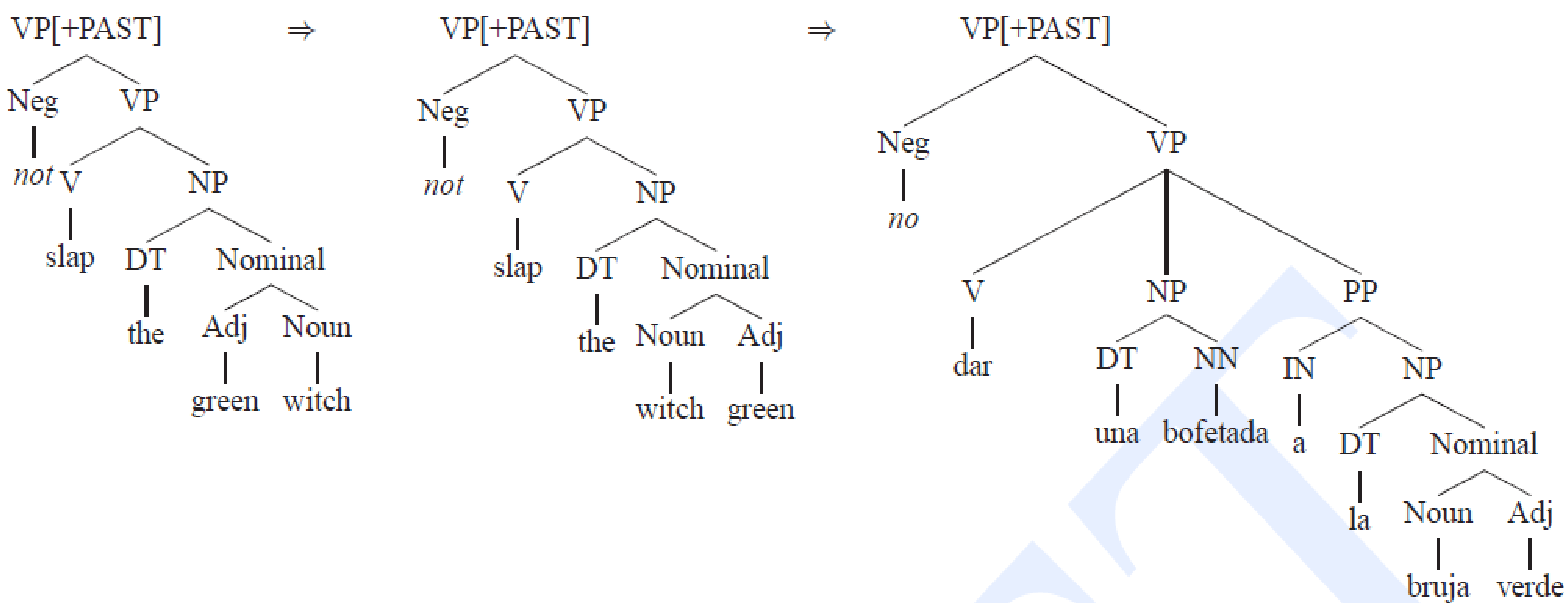
The green witch is at home this week.

Diese Woche ist die grüne Hexe zu Hause.

Transfer

- Use of constractive knowledge
  - Differences between the two languages
- Three phases
  - Analysis: parsing
  - Transfer
    - Lexical
      - English: home
      - German: nach Hause, Heim, Heimat, zu Hause, etc.
    - Syntactic:
      - Nominal (ADJ NN) ⇨ Nominal (NN ADJ)
  - Generation

Transfer



(figure taken from Jurafsky and Martin 2009)

Direct + Transfer

- Used by most commercial MT systems
- Shallow analysis
  - Morphological analysis, part-of-speech tagging
  - Chunking (NPs, PPs, etc.)
  - Shallow dependency parsing
- Transfer
  - Translation of idioms
  - Word sense disambiguation
- Syntesis
  - Lexical translation (bilingual dictionary)
  - Reordering
  - Morphological generation

Interlingual

- Shortcoming of transfer
  - Transfer rules for each pair of rules
  - Not feasible for multilingual environments (e.g., EU)
- Interlingual
  - Extract the meaning of the input and express it in the target language
  - Amount of knowledge transfer would be proportional to the number of language rather to the square
  - Deep semantic analysis from language X to the interlingual representation
  - Generating language Y from the interligual representation

Interlingual

- Semantic analyzer
  - „Mary didn't slap the green witch.“
    - EVENT: Slapping
    - AGENT: Mary
    - Tense: past
    - Polarity: Negative
    - Theme
      - Witch
      - Definiteness: DEF
      - Attributes
        - Has_color: green

Interlingual

- Advantages:
  - Natural language processing components can be used for both the translation and generation

- Shortcomings:
  - Semantic analysis is hard and only feasible for sublanguage domains
  - Unnecessary disambiguation across many languages
    - e.g., Chinese has concepts for ELDER_BROTHER and YOUNGER_BROTHER

Outline

- Task
- Classical MT
- Statistical MT
- Alignment in MT
- Evaluation

Statistical MT

- Model translation as the production of an output that maximizes a value function that represent the importance of faithfullness and fluency
- Probabilistic models of faithfullness and fluency
- Combine these models to choose the most probable translation for a sentence S:

$$best-translation\,\hat{T} = argmax_T\ faithfullness(T, S)\ fluency(T)$$

Statistical MT

- Foreign language sentence $F = f_1, f_2, \ldots, f_m$ to English
- The best English sentence $\hat{E} = e_1, e_2, \ldots, e_l$ is the one which maximizes P(E|F)

$$\hat{E} = argmax_E P(E|F)$$

$$\hat{E} = argmax_E \frac{P(F|E)P(E)}{P(F)}$$

$$\hat{E} = argmax_E \underbrace{P(F|E)}_{\text{translation model}} \underbrace{P(E)}_{\text{language model}}$$

Decoder: Given F and produces the most probable E

P(F|E): phrase-based translation model

- P(F|E): probability that sentence E generates sentence F
- Consider the behaviour of phrases
  - Which are usually moved as units
- Phrase-based translation model
  - Group source words into phrases: $\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_I$
  - Translate each source phrase $\bar{e}_i$ into a target phrase $\bar{f}_i$
  - (Optionally) reorder the target phrases $\bar{f}_i$

P(F|E): phrase-based translation model

- Translation probability: $\Phi(\bar{f}_i|\bar{e}_i)$
- Distortion probability (reordering):
  - distance between position of phrases in the two languages
  - a_i: start position of the target phrase generated by the i-th source phrase $\bar{e}_i$
  - b_{i-1}: end position of the target phrase generated by the (i-1)-th source phrase $\bar{e}_{i-1}$

$$d(a_i - b_i) = \alpha^{|a_i - b_{i-1} - 1|}$$

$$P(F|E) = \prod_{i=1}^{I} \phi(\bar{f}_i, \bar{e}_i)\, d(a_i - b_{i-1})$$

P(F|E): phrase-based translation model

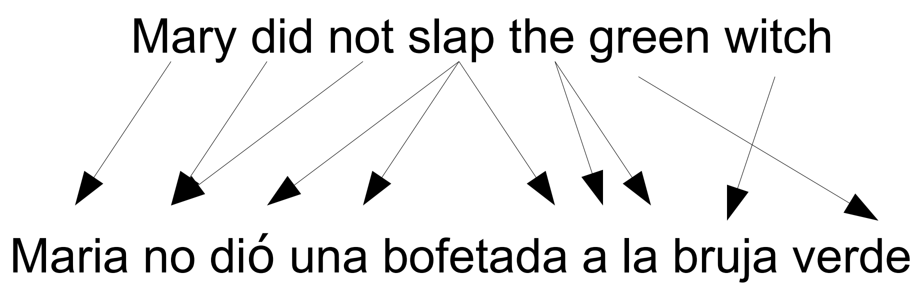
Position	1	2	3	4	5
English	Mary	did not	slap	the	green witch
Spanish	Maria	no	dió una bofetada	a la	bruja verde

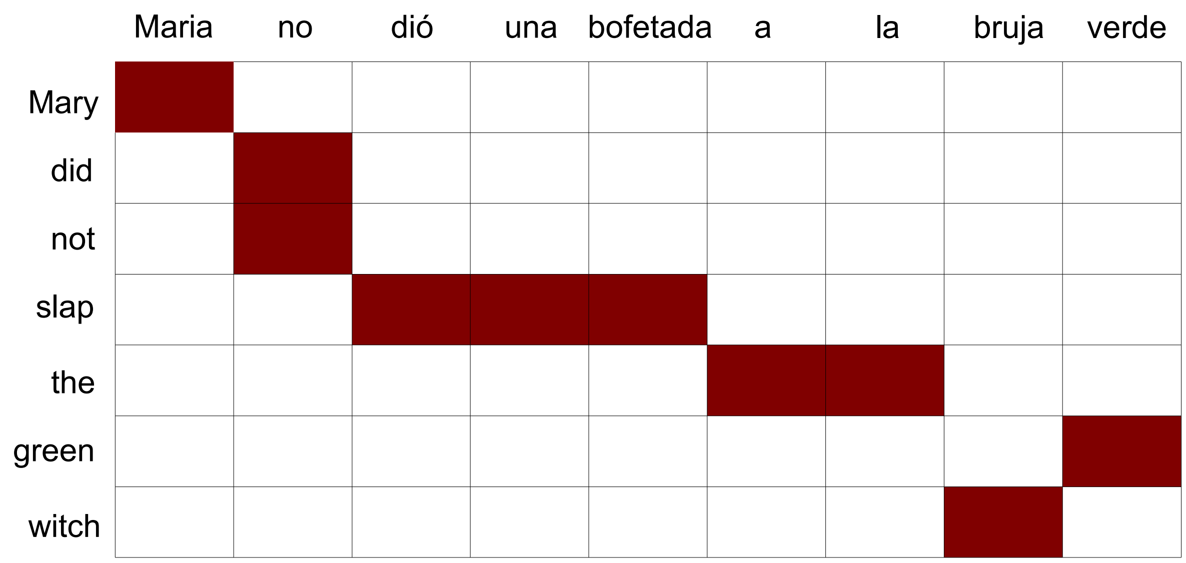
- Distortion are all 1
- P(F|E) = P(Maria,Mary) x d(1) x
  P(no|did not) x d(1) x
  P(dió una bofeatada|slap) x d(1) x
  P(a la|the) x d(1) x
  P(bruja verde|green witch) x d(1)

Statistical MT

- Translation probability: $\Phi(\bar{f}_i|\bar{e}_i)$
- Large bilingual training set
  - Phrase alignment: paired sentences and phrases
    - Difficult to get
  - Word alignment
    - From which we can get phrase alignments

Statistical MT





Outline

- Task
- Classical MT
- Statistical MT
- Alignment in MT
- Evaluation

Alignment in MT

- Word alignment:
  - Mapping source words to target words in a set of parallel sentences
- Each target word comes from exactly one source word
- Represent alignment as a sequence of index numbers

„And(1) the(2) program(3) has(4) been(5) implemented(6)“

„Le(1) programme(2) a(3) été(4) mis(5) en(6) application(7)“

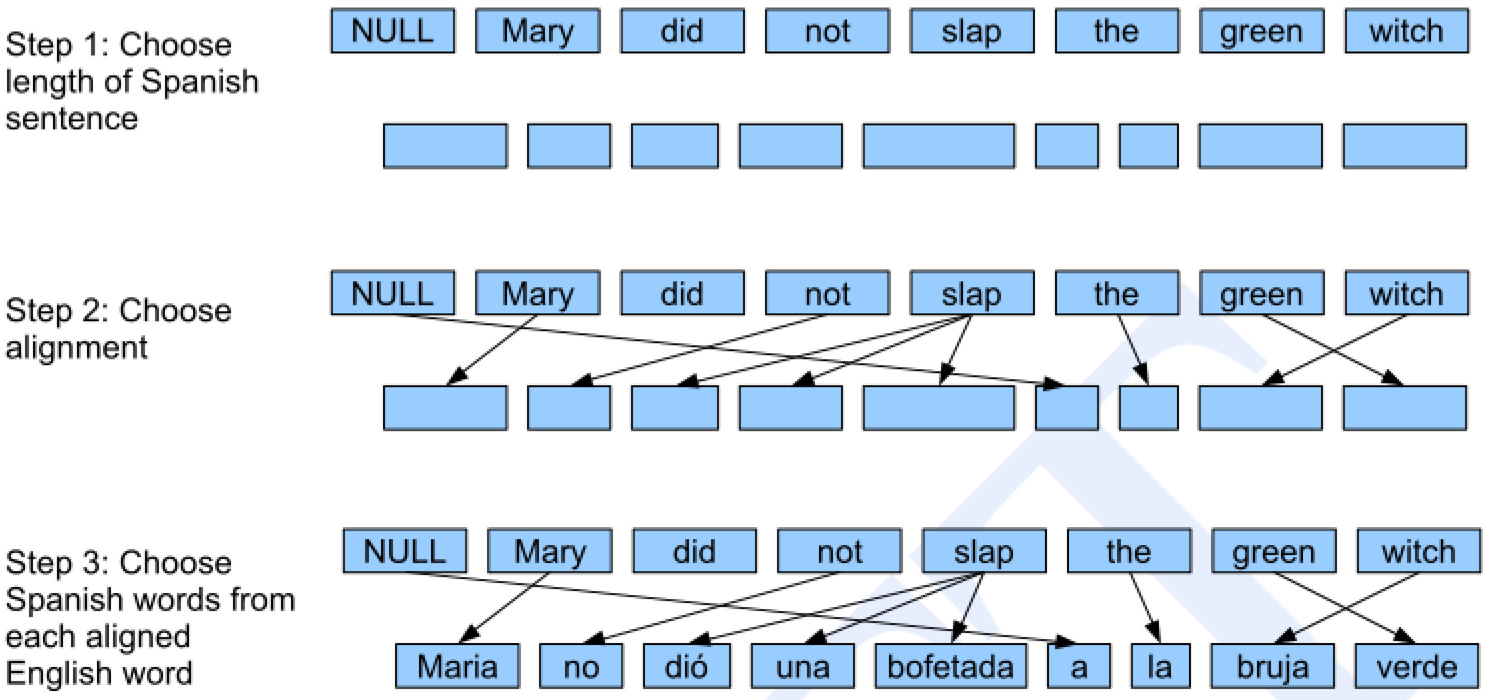
A = 2,3,4,5,6,6,6

IBM Model 1

- Statistical alignment algorithms

- We choose a length J for the target language $F = f_1, f_2, \ldots, f_J$
- Choose an alignment $A = a_1, a_2, \ldots, a_J$ between source and target languages
- For each position j in target sentence, choose a target word f_j by translating the source word aligned to it

IBM Model 1



(figure taken from Jurafsky and Martin 2009)

IBM Model 1

- e_{aj} is the source word that is aligned to the target word f_j
- $t(f_x,e_y)$ is the probability of translating e_y by f_x ($P(f_x|e_y)$)

$$P(F|E) = \sum_{A} \frac{\varepsilon}{(I+1)^J} \prod_{J=1}^{J} t(f_j | e_{a_j})$$

  - „I" English word (+1 NULL)
  - Small constant ε: probbaility of choosing length J

- Very simplifying assumption: all alignments are equally likely
- Decoding algorthm (Viterbi) and training the model (EM)

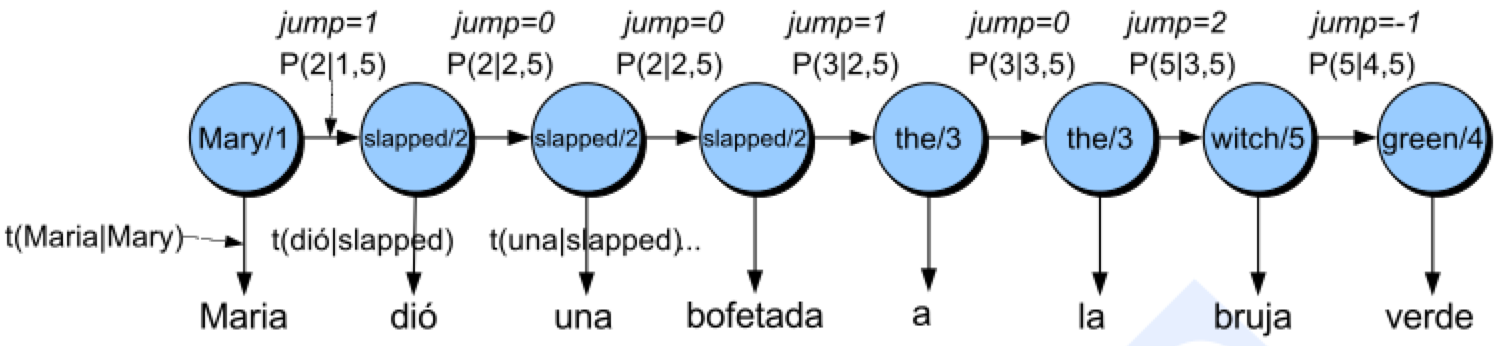
HMM alignments

- Not all alignments are equally likely
- Alignments tends to preserve locality
- HMM alignment: restructure the P(F,A|E) using the chain rule

$$P(f_1^J|e_1^J) = P(J|I) \times \sum_A \prod_{j=1}^{J} P(a_j|a_{j-1}, I) P(f_j|e_{a_j})$$

  - The probability of a_j depends only on previous aligned a_{j-1}
  - The probability of f_j depends only on the aligned word e_{a_j} at position j
  - Length probability is P(J|I)

HMM alignments



P(F,A|E) = P(J|I) x
P(Maria,Mary) x P(2|1,5) x
P(dió|slapped) x P(2|2,5) x
P(una|slapped) x P(2|2,5) x ...

(figure taken from Jurafsky and Martin 2009)

Training alignment models

- Parallel corpus (parallel text, bitext)
  - Canadian Parlament (Hansards): English, French
  - Hong Kong Hansards: English, Chinese
  - United Nations
- Literary translation
  - License problems
  - No literal translation
- Bible
- Wikipedia

Training alignment models

- Training
  - Sentence segmentation
  - Sentence (word) alingment
  - Sentence which can't be aligned are thrown away

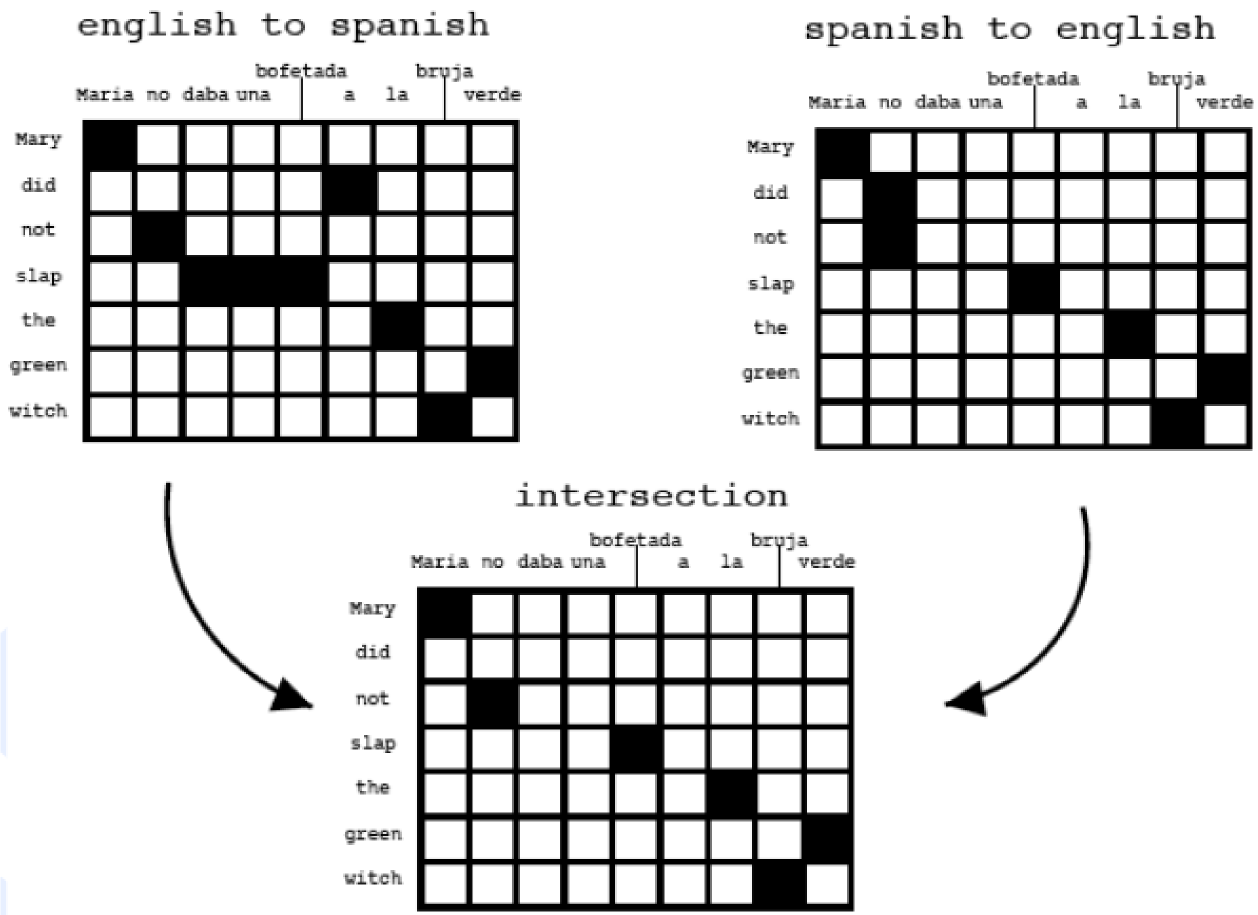
  - Input: sentence pairs $\{(F_s,E_s): s=1...S\}$
  - Goal: learn alignment $A = a_1^J$ and the probabilities (t for Model 1, lexicon and alignment probabilities for HMM)
  - EM algorithm

Symmetrizing Alignments for Phrase-Based MT

- Production of phrase-to-phrase alignments

- Start taking the intersection of the two alignments (high precision)
- Compute separately the union of the two alignments (less acurate)
- Build a classifier to select words from the union and add them to the intersection
- Calculate probabilities for the various alignments

Symmetrizing Alignments for Phrase-Based MT



(figure taken from Jurafsky and Martin 2009)

Symmetrizing Alignments for Phrase-Based MT

(Maria, Mary), (no, did not),
(slap, dió una bofetada), (verde, green),
(a la, the), (bruja, witch),
(Maria no, Mary did not),
(no dió una bofetada, did not slap),
(dió una bofetada a la, slap the),
(bruja verde, green witch),
(a la bruja verde, the green witch),. . .

Decoding for Phrase-Based Statistical MT

- Take a source sentence and produce the best target translation according to the product of translation and language models

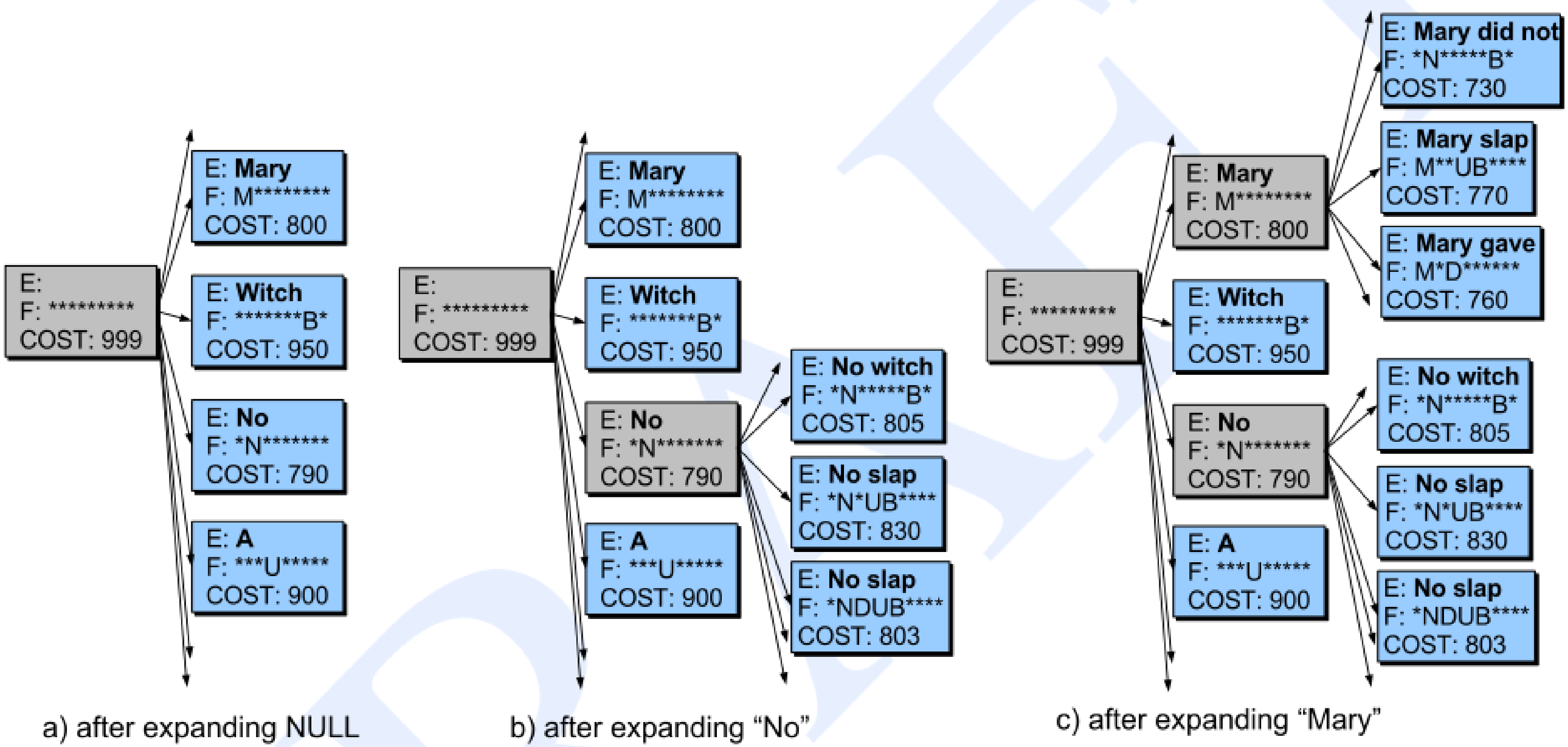
$$\hat{E} = argmax_E \, P(E|F) \, P(E)$$

- Search problem
- Best-first search (heuristic, informed search)
  - Do not consider all target sentences
  - But only those which are possible translations

Decoding for Phrase-Based Statistical MT

- Algorithms
  - Stack decoding
    - Priority queue
    - Current cost
      - Total probability of phrases which have been translated so far (translation, distortion, language model)
    - Future cost
      - Total probability of remaining search path (w/o distortion)
  - Beam-search pruning
    - Pruning high-cost states (keep best n entries)

Example of Stack decoding



(figure taken from Jurafsky and Martin 2009)

Outline

- Task
- Classical MT
- Statistical MT
- Alignment in MT
- Evaluation

Evaluation

- Two dimensions: fidelity and fluency
- Fluency
  - Scale (human raters): 0 (totally unintelligible) to 5 (totally intelligible)
  - Measure of time (fluent text are read faster)
- Fidelity
  - Adequacy
    - Bilingual raters (compare both texts)
    - Monolingual raters (good gold standard)
  - Informativeness
    - Multiple choice test about the text

Evaluation

- Automatic methods
  - Bleu, NIST, TER, Precision, Recall, METEOR
- Comparison to multiple human translations
  - Good translations might look different
- Bleu
  - Rank the MT output by a weighted average of the number of N-grams overlaps with the human translations

Bleu metric

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

(figures taken from Papineni et al. 2002)

Bleu metric

- Unigram:
  - Candidate length = 10
  - Matches = 6
  - 6/10 = 0.6
- Example
  - the the the the the the the
  - the cat is on the mat
  - Bleu = 7/7 = 1.0
  - Modified N-gram precision: Bleu = 2/7
  - Limited by the number of times a word appear in the reference

Bleu metric

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Modified P (unigram) = 17/18
Modified P (bigram) = 10/17

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Modified P (unigram) = 8/14
Modified P (bigram) = 1/13

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

(figures taken from Papineni et al. 2002)

Bleu metric

- Modified N-gram precision:

$$p_n = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in Candidates} \sum_{n-gram' \in C'} Count(n-gram')}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases}$$

$$Bleu = BP \times \exp\left(\frac{1}{N} \sum_{n=1}^{N} \log p_n\right)$$

(Effective reference length *r*   BP=Brevity penalty)

Further Reading

- Speech and Language Processing
  - Chapters 25