Lecture 26: MT, Formal Language Theory (1)

Ling 1330/2330 Intro to Computational Linguistics

Na-Rae Han, 12/5/2023

Overview

- Machine translation
  - Noisy channel model
  - Alignment, parallel corpora
  - Neural MT

- Formal language theory
  - Eisenstein (2019) Ch.9 Formal language theory, draft copy
  - *Mathematical Methods in Linguistics* by B. Partee, A. ter Meulen and R. Wall
    - Excerpt posted on Canvas, under "Modules"

Statistical MT

- The three classic architectures focus on the appropriate representations to use → symbolic approach
- In **statistical MT**, however, the focus is more on the result
  - JAPANESE: *fukaku hansei shite orimasu.*
  - English1: *I sincerely apologize.*
  - English2: *I am deeply regretting.*

  ← E1 is more **natural** English; E2 is more **faithful** to the original meaning

  ← There are two considerations in translation:

  (1) **Faithfulness** to the original message

  (2) **Fluency (naturalness)** of the target language text

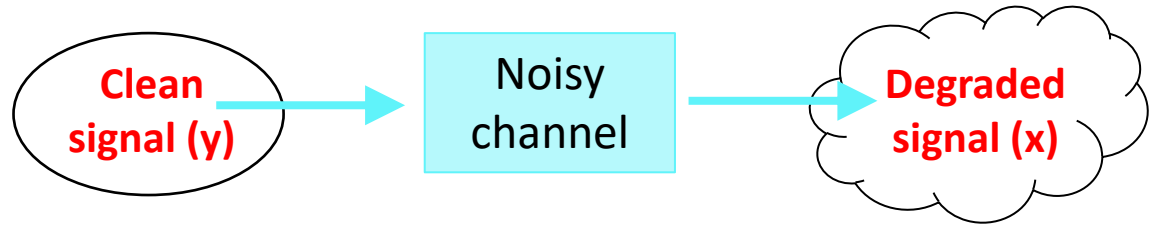
  ← Successful translation can be schematized as:

  A translation T that maximizes the product of the two

  best-translation $\hat{T} = \text{argmax}_T$ faithfulness(T, S) * fluency(T)

The noisy channel model

- The **noisy channel model**



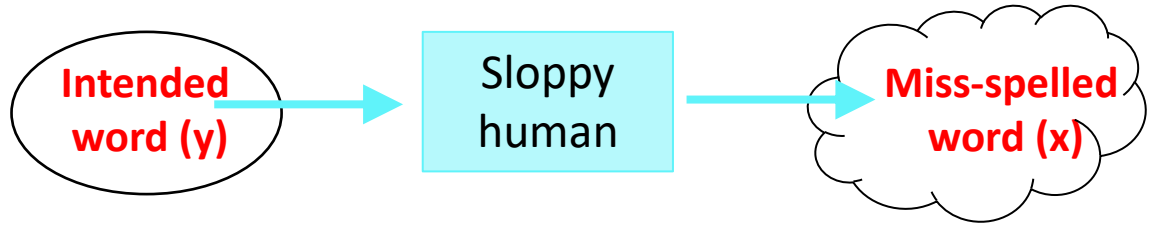
$$\hat{y} = \text{argmax}_y P(y)P(x \mid y)$$

- How to estimate the original clean signal from the degraded signal?
  - ŷ: system's best guess what the original message was
  - P(y): the probability of the original message
  - P(x|y): the observed output of the channel, given y

  ← ŷ is the message that maximizes P(y) * P(x|y)
- Became a powerful conceptual frame for a variety NLP tasks!

Spell correction as noisy channel model

- Spelling error correction

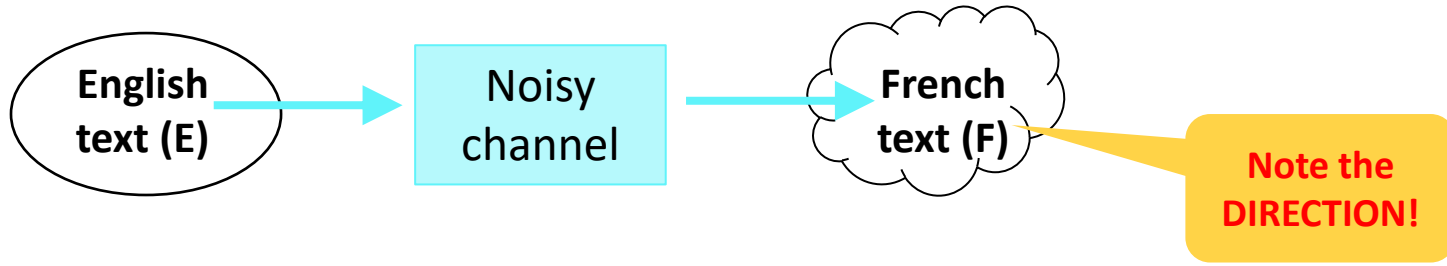


$$\hat{y} = \text{argmax}_y P(y)P(x \mid y)$$

- Estimating the intended word based on the mis-spelled word
  - ŷ: system's best guess what the intended word was, e.g., *bright*?, *birth*? *broth*?
  - P(y): the probability of the intended word, e.g., P(birth) vs. P(broth)
  - P(x|y): the probability of the misspelled word, given the intended word, e.g., P(briht|birth), P(briht|broth), ...

  ← ŷ is the word that maximizes P(y) * P(x|y)

Translation as noisy channel model



- Translation model, French (F) to English (E):

$$\hat{E} = \text{argmax}_E P(E \mid F) = \text{argmax}_E \boxed{P(E)} \boxed{P(F \mid E)}$$

Language model **Translation model**

  - Ê: system's best guess at best English translation
  - P(E): the probability of the English translation text
  - P(F|E): the observed distribution of French text, given the English text
- We need two components:
  The English language model P(E) and **the translation model P(F|E)**

  ← Based on the two, a **decoder** picks $\hat{E}$, i.e., determines the best English translation

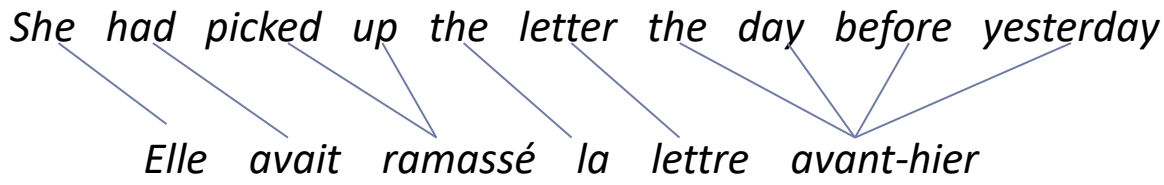
P(E) vs. P(F|E)

- Two components:
  - **The English language model:** P(E)
    - ←Essentially a model of **fluency/naturalness** ("I sincerely apologize" more likely English sentence than "I am deeply regretting")
    - ←We can build this independently based on a large English corpus

      (Bigram model, trigram model, HMM, ...)
  - **The translation model:** P(F|E)
    - ←Essentially a model of **translation faithfulness** ("I am deeply regretting" more faithful to original Japanese than "I sincerely apologize")
    - ←How should we obtain this?

      Word alignment from a **parallel corpus**
    - ←**As with any other statistical language models, this model is REDUCTIVE: the stats come from word-level co-occurrence patterns**

Word alignment



- The translation model P(F|E) based on word alignment statistics
  - Sentences are reduced to a bag-of-words
  - **Given an English sentence as words (w_e1... w_en), how likely is it to see a French sentence (w_f1... w_fn) with a set of alignment (A) between words?**

$$P(F \mid E) = \sum_{A} P(F, A \mid E)$$

- Where to get **word-level alignments (A)**?

  ← **It can be automatically extracted from a large parallel corpus**, where sentences are pre-aligned.

  ← There are many methods for extracting word alignment. (IBM Model 1, 2, 3, 4; HMM, ...)

Parallel corpora

- **European Parliament Proceedings Parallel Corpus 1996-2011**
  - **Europarl**: http://www.statmt.org/europarl/
  - Extracted from the proceedings of the European Parliament.
  - Includes versions in 21 European languages. (EU currently has 23 official languages)
  - Each language section is aligned on sentence level
- EU and MT
  - Public official documents of the European Union (EU) must be translated to every official languages
  - EU has been a big sponsor for MT research efforts.

Europarl Corpus in NLTK

```
>>> from nltk.corpus.europarl_raw import english, german, french, spanish
>>> english.sents()[2]
['You', 'have', 'requested', 'a', 'debate', 'on', 'this', 'subject', 'in', 'the',
'course', 'of', 'the', 'next', 'few', 'days', ',', 'during', 'this', 'part-session',
'.']
>>> german.sents()[2]
['Doch', 'sind', 'Bürger', 'einiger', 'unserer', 'Mitgliedstaaten', 'Opfer', 'von',
'schrecklichen', 'Naturkatastrophen', 'geworden', '.']
>>> french.sents()[2]
['En', 'revanche', ',', 'les', 'citoyens', "d'", 'un', 'certain', 'nombre', 'de',
'nos', 'pays', 'ont', 'été', 'victimes', 'de', 'catastrophes', 'naturelles', 'qui',
'ont', 'vraiment', 'été', 'terribles', '.']
>>> spanish.sents()[2]
['En', 'cambio', ',', 'los', 'ciudadanos', 'de', 'varios', 'de', 'nuestros',
'países', 'han', 'sido', 'víctimas', 'de', 'catástrofes', 'naturales',
'verdaderamente', 'terribles', '.']
```

Also available: Danish, Dutch, Finnish, Greek, Italian, Portuguese, Spanish, Swedish

Parallel corpus with syntactic annotation

```
;;05:127: 저는 그 일을 할 수 있는 한 빨리 하겠습니다 .

(S (NP-SBJ 저/NPN+는/PAU)
   (VP (NP-OBJ-LV 그/DAN
                  일/NNC+을/PCA)
       (VP (NP-ADV (S (NP-SBJ (S (NP-SBJ *pro*)
                                 (VP 하/VV+ㄹ/EAN))
                              (NP 수/NNX))
                      (ADJP 있/VJ+는/EAN))
                  (NP 한/NNX))
           (ADVP 빨리/ADV)
           (VP (LV 하/VV+겠/EPF+습니다/EFN))))
   ./SFN)
```

```
;;05:127: I will do it as fast as possible.

(S (NP-SBJ (PRP I))
   (VP (MD will)
       (VP (VP (VB do)
               (NP-OBJ (PRP it)))
           (ADVP-MNR (ADVP (RB as)
                           (RB fast))
                     (PP (IN as)
                         (ADJP (JJ possible))))))
   (. .))
```

Korean English Treebank Annotations: https://catalog.ldc.upenn.edu/LDC2002T26

A tectonic shift: Neural MT

SMT versus NMT

- **Statistical machine translation**: decoder takes into account information from multiple models (translation model, language model, reordering model) to generate a translation hypothesis
- **Neural machine translation**: a vector representation in one language is used to find a similar representation in another language
- Many more components in statistical MT, more moving parts but it's easier to see what is happening
- Neural MT is less complex, but it's much harder to understand what is going on and how a particular translation is generated

SMT versus NMT

NMT is State-of-the Art

- Fluent!
- Accurate
- Better performance on unrelated languages But…
- Requires LOTS of data
- Can be very unpredictable

What about SMT…?

- Requires much less data
- Easily interpretable and predictable, no surprises
- Robust to noise and variation in input
- Better at short segments
- Better at certain downstream tasks (Information Retrieval)

Neural MT: circling back

More details in:

- Tianyi's presentation "Deep Learning Language Models"
- "The Great AI Awakening" (NYT, HW10 reading)

Formal Language Theory

RE/FSA and natural language grammar

- RE/FSA can be thought of as a *grammar*.
  - RE/FSA is a mechanism that rules if a string is accepted ("grammatical") or not ("ungrammatical")
  - An RE/FSA represents a set of accepted strings (= "grammatical strings") to the exclusion of the rest ("ungrammatical strings")
  - A *language* is a set of all grammatical strings generated over its vocabulary (arc labels).
  - *L*(*ab*a*) = a language generated by regular expression *ab*a*

    = {*aa*, *aba*, *abba*, *abbba*, *abbbba*, *abbbbba*, ... }

English morpho-syntax as FSA

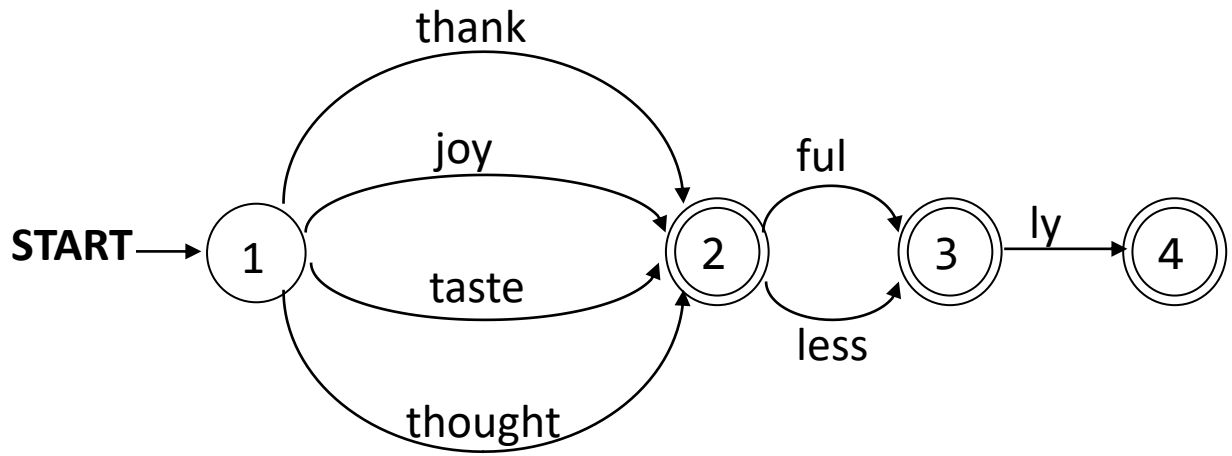
- A set of English morphemes:
  - {thank, joy, taste, thought, ful, less, ly}

- You can concatenate the morphemes to make English words.
  - **Grammatical**: joy, joyful, joyless, joyfully, …
  - **Ungrammatical**: ful, fuljoy, lessful, joythank, joylessthankthank, …

- English morpho-syntax dictates which sequences are grammatical English words and which are not.
  - How to formalize the rules?

English morpho-syntax as FSA

- Arc labels (=vocabulary): *English morphemes*
- Set of accepted strings: *legitimate English words*



- Which words are in this language?
- Which are not?
- What's the corresponding regular expression?
  - (thank|joy|taste|thought)((ful|less)(ly)?)?

English phonotactics as FSA

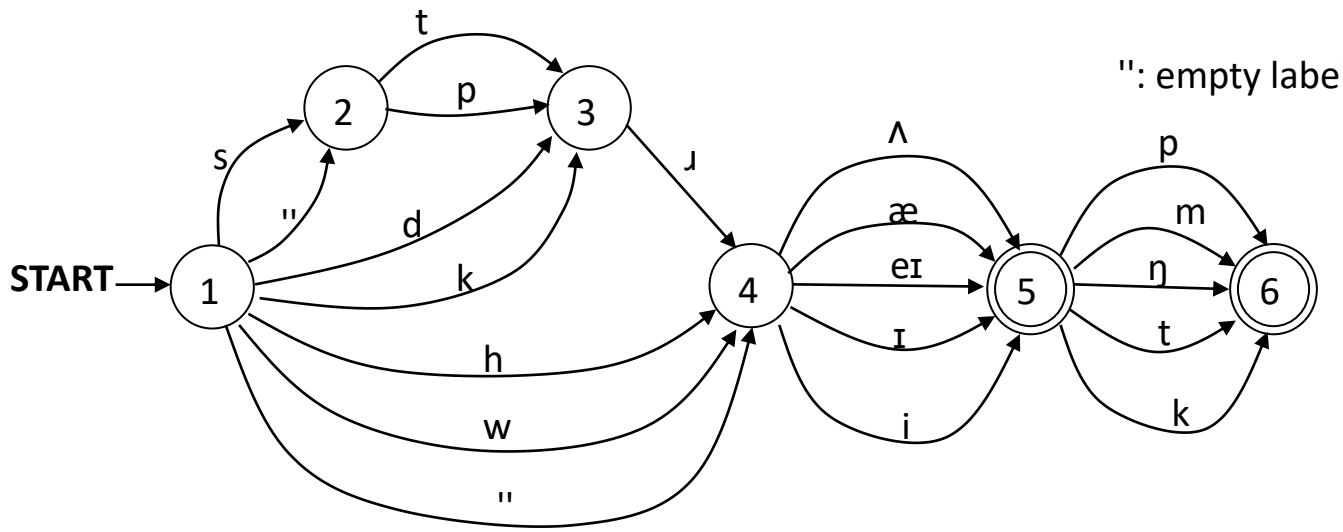
- A set of English phonemes:
  - {g, k, b, p, d, t, s, z, v, f, θ, ð, ʃ, ʒ, tʃ, dʒ, ɹ, l, h, m, n, ŋ, j, w, a, æ, ɛ, i, ɪ, ɔ, ʌ, ɑ, u, ʊ, eɪ, oɪ, aʊ, ou, aɪ}

- You can concatenate the phonemes to make English syllables.
  - **Grammatical**: pɔ, deɪt, æbz, stɹid, …
  - **Ungrammatical**: mk, gki, iu, mɑnpk, ɹstid, …

- English phonotactics dictates which sequences are grammatical English syllables and which are not.
  - How to formalize the rules?

English phonotactics as FSA

- Arc labels (=vocabulary): *English phonemes*
- Set of accepted strings: *valid English syllables*



- Which syllables are in this language?
- Which are not?
- What's the corresponding regular expression?

English syntax as FSA

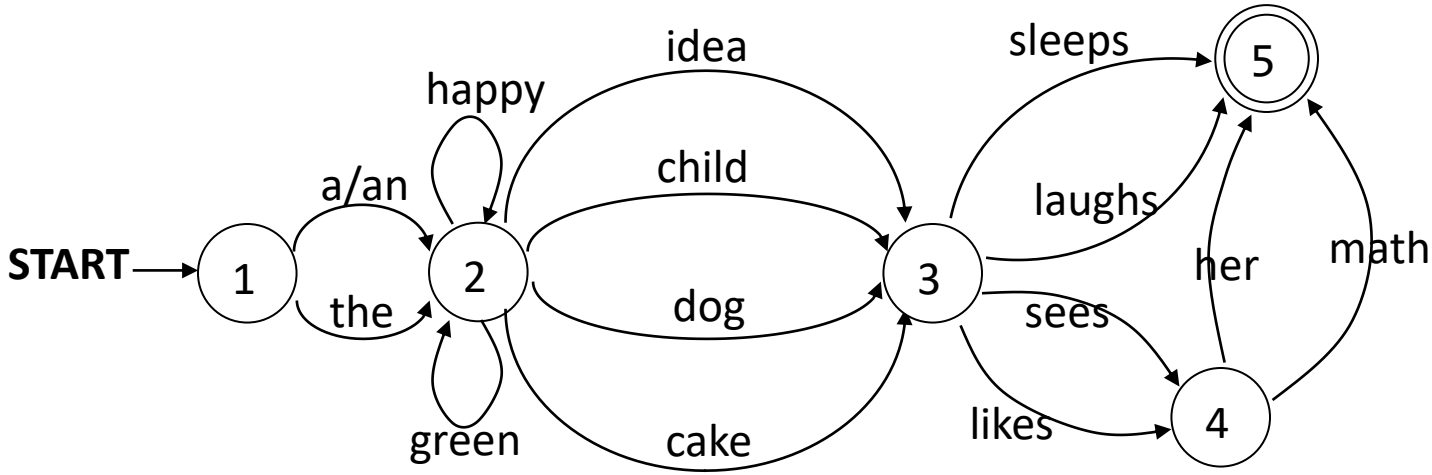
- A set of English words:
  - {a/an, the, happy, green, child, dog, idea, cake, sleeps, laughs, sees, likes, her, math}

- You can concatenate the words to make English sentences.
  - **Grammatical**: a green dog likes the idea
    the happy cake sleeps
    a dog sees math
  - **Ungrammatical**: dog a green the likes idea
    cake dog her
    laughs dog a the

- English syntax dictates which sequences are grammatical English sentences and which are not.
  - How to formalize the rules?

English syntax as FSA

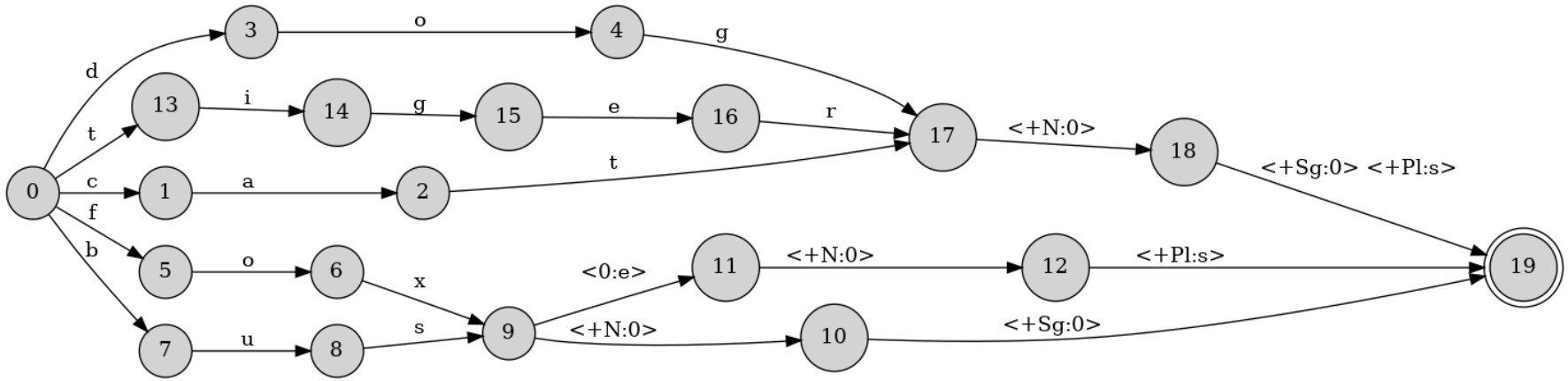
- Arc labels (=vocabulary): *English words*
- Set of accepted strings: *grammatical sentences*



- Which sentences are in this language?
- Which sentences are not?
- What is the corresponding regular expression?

FST: a variant of FSA

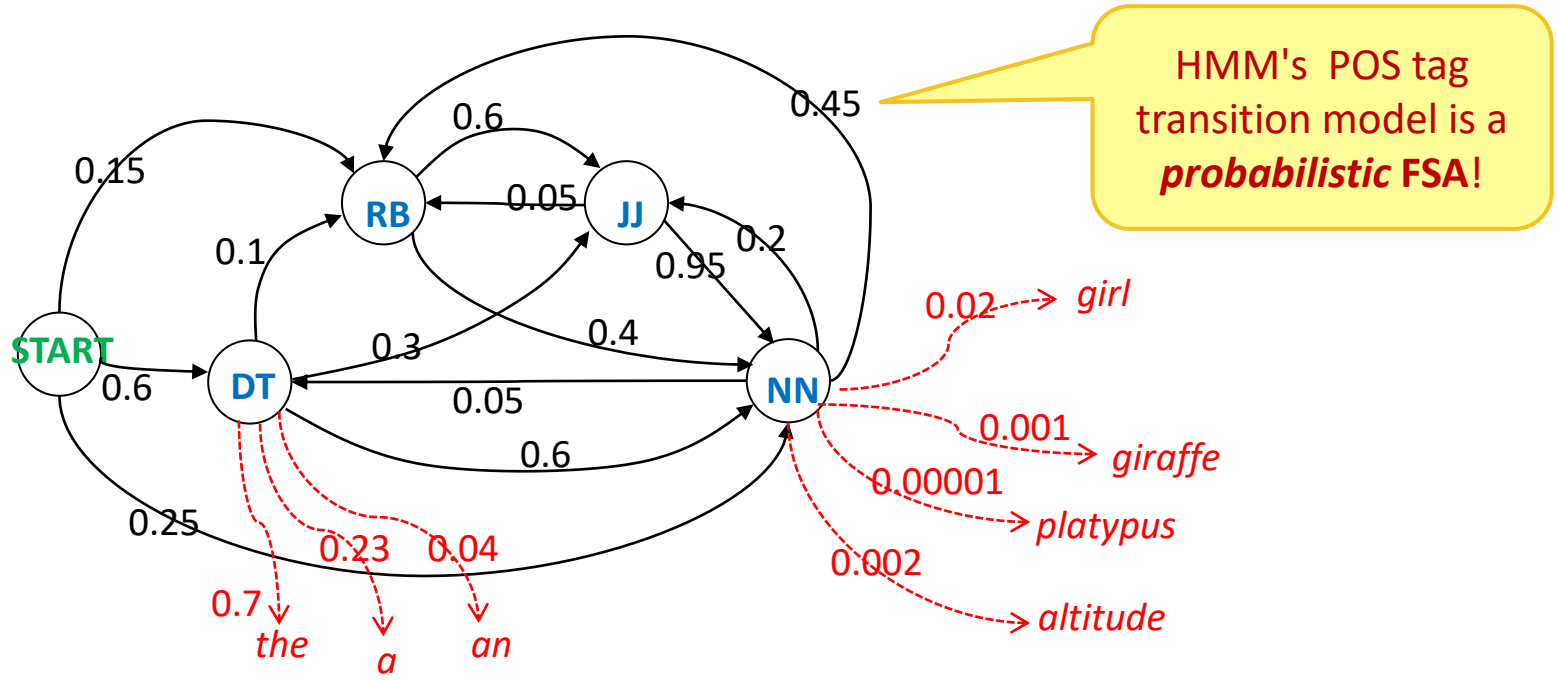
- Finite-state transducers are a FSA with two-sided arc labels.



- Models morpho-syntax (FSA) as well as morphological analysis/generation (transduction).

HMM builds on FSA

- HMM combines POS tag sequence probability (DT → JJ → NN → …) and the probability of certain words occurring with a POS (given DT, 'the' is 0.7 likely, and 'a' 0.23…)



Are FSA good enough?

Question:

- Is the Finite-State Machine *powerful enough* to capture the grammatical system of English phonology?
- How about English morpho-syntax?
- How about English syntax?

⇐This inquiry forms the basis of the **formal language theory**.

A formal definition of *language*

- Alphabet (vocabulary) = A = {a, b}
- The largest possible *language* generated on A:

e: an empty string (= '')

$L_0 = A^* = \{e, a, b, ab, ba, aab, bba, aba, bab, \ldots aabbaaababbaa, \ldots\}$

← Any string that results from concatenation of {a, b} is in this language, i.e., grammatical. There is no ungrammatical string.

← This is an infinite set.

- A **language** over vocabulary A is *any subset* of A*.

L_1 = {x | x contains any number of a's followed by a single b.}

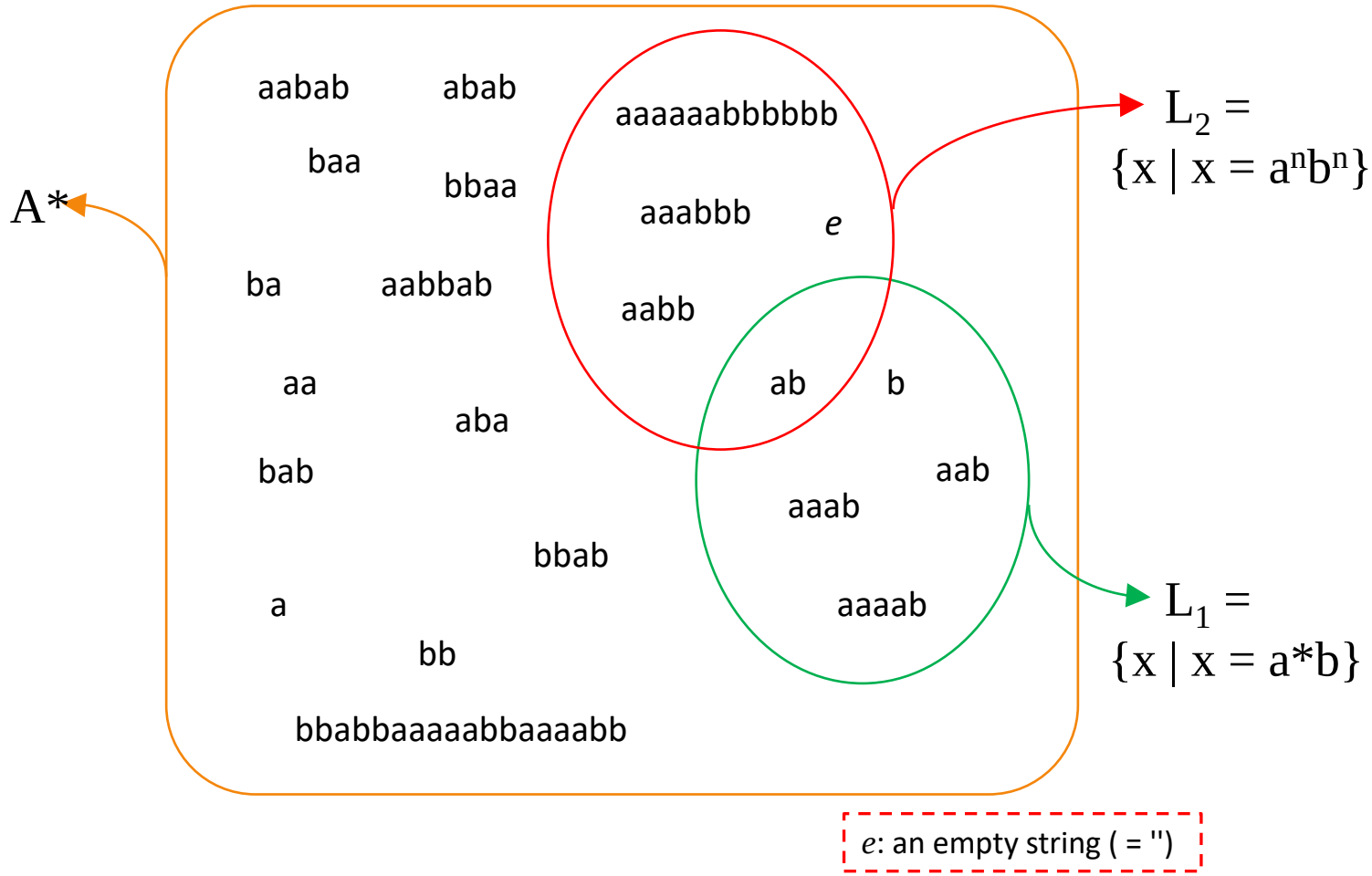
= {b, ab, aab, aaab, aaaaab, ..., aaaaaaaaaab, ...}

← These are grammatical strings for L_1

- Strings that are *not* in this language:

e, a, aa, aba, aabbb, ... ← These are ungrammatical strings

Languages made out of a's and b's



Small alphabet, *a lot* of languages

Alphabet (vocabulary) = A = {a, b}
A* = {*e*, a, b, ab, ba, aab, bba, aba, bab, ... aabbaaabababbaa, ...}
A *language* over A is *any* *subset* of A*.

- How many different languages are there?
  - An infinite number (the power of the size of integers): $2^{\aleph_0}$
- Examples:

  L_1 = {x | x is 2 characters long or shorter} = {*e*, a, b, aa, bb, ab, ba}

  L_2 = {x | x contains any number of a's followed by a single b}
  = {b, ab, aab, aaab, aaaaab, ..., aaaaaaaaaab, ...}

  L_3 = {x | x contains an even number of a's}

  L_4 = {x | x has form $a^n b^n$; some # of a's followed by the same # of b's}

  L_5 = {x | x contains equal numbers of a's and b's in any order}

Are all languages *equally complex*?

- Languages over A = {a, b}:
  - L_1 = {x | x is 2 characters long or shorter}
  - L_2 = {x | x contains any number of a's followed by a single b}
  - L_3 = {x | x contains an even number of a's}
  - L_4 = {x | x has form $a^n b^n$}
  - L_5 = {x | x contains equal numbers of a's and b's in any order}
  - L_6 = {x | x is a palindrome}
  - L_7 = {x | x has form *ww*, i.e., consists of two halves that are identical}
  - L_8 = {x | x contains #-many a's where # is a prime number}

"copy" language

- Questions:
  - Are some languages *more complex* than others?
  - Which languages are on the *same complexity scale level*?

Wrapping up

- Thursday class:
  - Formal language theory
  - Course wrap

- Homework 10 due on Thu
  - Submit on Canvas
  - Also: **extra credit** opportunity, on MS Teams

- Grades, late work forgiveness ➔
- Extra credit ➔
- Final exam info ➔

Your grade: what's ahead

- **Canvas's Grade Center** is being prepped
  - Your exercise score is in
  - Homework 9 and 10 grades are outstanding
  - Attendance & participation records (will post 2nd half attendance soon)
    - **1 missed class exemption → raised to 2**
  - Weighted running total (CAVEAT!!)

- **Late work forgiveness**
  - Everyone gets one make-up opportunity. Choose from:
    1. Finish up an incomplete homework submission or re-do a part, no penalty.
    2. Up to 3 days of late submission penalty waived.
    3. Missed homework: 25% penalty. Upload on Canvas and email me.
    4. Missed exercise: 5/10 for satisfactory (80+%) work. Email me as attachment.
  - Deadline: **12/15 (Fri) 11:59pm**. Email me and let me know of your choice!
  - If a solution has been published, feel free to look it up. It's fine as long as you don't blindly copy it. (Make sure to demonstrate you are not blindly copying.) There's already a late penalty, and I'd rather you learn.

Extra credit, round-up

- If you have 100% on Exercises, you are already eligible for a standard round-up, up to 0.4%.
  - Normally 89.6% is B+; it will be bumped up to A- (90%)

- Extra credit opportunity (1): NLP talk
  - Attend Linguistics dept colloquium:
    - Dec 1 (Fri) 3pm, G8 CL. Lorraine Li, "Probabilistic (Commonsense) Knowledge in Language"
    - If you can't attend this one, find a different CL/NLP talk (CMU, Pitt, online)
  - Submit a short report on Canvas, earn 0.3% extra credit

Both due 12/15 (Friday) 11:59pm

- Extra credit opportunity (2): share HW10 essays
  - On MS Teams, share your HW 10 essay, read 3 classmates' essays and leave comments, earn 0.3% extra credit

Final exam

- 12/13 (Wed), 4—5:50pm
- At G17 CL (Language Media Center)

- 150 total points (50% larger than midterm)
- All pen-and-pencil based.
- **1 cheat sheet allowed**:
  - letter-sized, front-and-back, hand-written.
- Cumulative! 10-20% will be from first half of the semester.
- Make sure to study book chapters and other linked materials. Post-midterm, my slides are not as "comprehensive".