



COURSE RESOURCES

Reminder



[Course homepage](#)

LECTURE 4: MEANING, REPRESENTATIONS AND FINITE-STATE METHODS

Mark Granroth-Wilding

Get from here to:

- [Assignment instructions](#)
- [Moodle](#): submit assignments, discussion forum
- Lecture slides: incl. suggested reading
- Links to reading material
- [Presemo](#): ask questions in lectures

1

ANNOTATION

- Sometimes need new annotated corpus
- Some things to decide on to estimate **cost** (time and money):
 - Source data: genre, size, licensing
 - Annotation scheme: complexity, guidelines
 - Annotators: expertise, training
 - Annotation software: graphical?
 - Quality control: multiple annotation? adjudication?
- Competent annotator: some tasks are straightforward for most inputs
- Others not:
 - Ambiguous text
 - Grey areas between categories

Based on slides by Henry Thompson, Edinburgh University, with permission

2

YOU PLAY ANNOTATOR

Verb, noun or adjective?

- *We had been **walking** quite briskly*
- ***Walking** was the remedy, they decided*
- *Sandburg was a **walking** thesaurus of American folk music*
- *We all lived within **walking** distance of the studio*
- *A woman came along carrying an umbrella as a **walking** stick*
- *The Walking Dead premiered in the US on Oct 31 2010*

Based on slides by Henry Thompson, Edinburgh University, with permission

3

ANNOTATION: NOT AS EASY AS YOU THINK

- Any annotation scheme has some **difficult cases**
- Grey areas, multiple plausible decisions
- Human language is flexible: cuts corners, changes over time
- Seen some examples already (POS tags, syntax)
Much more in lecture 14
- Want annotations to be **clean, consistent, reliable**
- **Annotation manual**: important for producing consistent data and making annotators' job easier

Based on slides by Henry Thompson, Edinburgh University, with permission

4

ANNOTATION MANUAL: PTB

- Penn Treebank: 36 POS tags
- Tagging guidelines: 34 pages
'The temporal expressions yesterday, today and tomorrow should be tagged as nouns (NN) rather than as adverbs (RB). Note that you can (marginally) pluralize them and that they allow a possessive form, both of which true adverbs do not.'
- Tree bracketing guidelines: >300 pages!

Based on slides by Henry Thompson, Edinburgh University, with permission

5

ANNOTATION QUALITY

Even with good guidelines, annotations won't be perfect

- Simple errors (wrong button)
- Not reading full context
- Forgetting detail from guidelines
- Cases not anticipated by guidelines

How to measure quality of 'gold' data?

Based on slides by Henry Thompson, Edinburgh University, with permission

6

MEASURING QUALITY

- Multiple people independently annotate same data
- Measure **inter-annotator agreement** (IAA)
- **Raw agreement rate**: proportion of labels in agreement
- If low, examine disagreement
 - More training
 - Refine guidelines
 - Reject some data
- More sophisticated measures account for:
 - Knowledge of annotation scheme (some things harder / more important)
 - Probability of agreement by chance
- **Upper bound** (*human ceiling*) for system accuracy on task

Based on slides by Henry Thompson, Edinburgh University, with permission

7

OVERFITTING ON A GRAND SCALE

- Another of overfitting: **long-term**
- Standard test set T
 - Everyone evaluates on T
 - All observe proper practices, keep T blind
 - Better models (on T) 'win', built on in further work
 - After 20 years, SOTA models excellent at T 's
 - annotations
 - annotation type
 - task
 - domain
 - language
 - But no good outside these contexts
- Same applies to **evaluation metrics** and **methods**

8

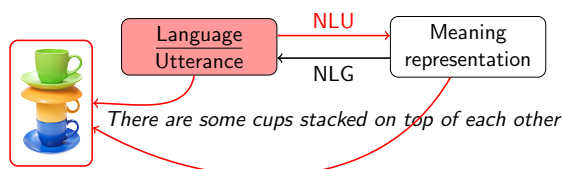
ANNOTATION

Summary

- Considerations when building a corpus:
 - Data, annotation scheme, annotators, quality control
- Annotation is hard!
 - Grey areas
 - Ambiguity, lack of clarity in language
 - Want clean, reliable annotations
- Guidelines / manual
- Measuring quality
- Long-term overfitting

10

TWO PERSPECTIVES



Dual interpretation of meaning representation:

Meaning of utterance

State of the world
(described by utterance)

We'll come back to this...

12

WHAT TO REPRESENT?

Example

The Kenora Thistles were an ice hockey team.

- There was an entity (team)
- It was called *The Kenora Thistles*
- It was an ice hockey team

14

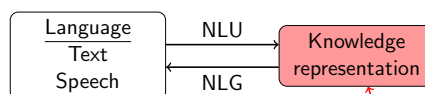
OVERFITTING ON A GRAND SCALE

What can we do?

- Do improvements help with downstream applications?
 - **Extrinsic** evaluation
- No **metric** tells the full story
- Compare models on a range of
 - tasks
 - test sets (domains, languages, text types, ...)
 - metrics
- Error analysis

9

NATURAL LANGUAGE PROCESSING



Natural Language **Understanding** (NLU)
Natural Language **Generation** (NLG)

- Knowledge/meaning representation: depends on application
- What type of meaning?
- How to represent formally?
- How to use computationally? Inference, reasoning, ...

11

WHAT TO REPRESENT?

Meaning of utterance \leftrightarrow Knowledge of world

- Assume meaning of expression closely related to **knowledge**
- Representations contain *same kind of stuff*

Example

I have a car.

- There is some car
- The speaker owns it

13

EXERCISE: WHAT TO REPRESENT?

In small groups

- **Task:** Automatic restaurant suggestion
- **Input:**

Example

I like noisy, expensive restaurants that serve vegetarian dishes.

- List elements of input's meaning relevant to task
- Suggest some ways to represent these
- Don't worry about *how* to extract them!

15

EXERCISE: WHAT TO REPRESENT?

- **Task:** Modelling the described scene
- **Input:**

Example

*In one corner of the room was a small, darkish-brown mouse.
Opposite it, to my left, was a large creature, unfamiliar to me.*

- List elements of input's meaning relevant to task
- Suggest some ways to represent these
- Don't worry about *how* to extract them!

16

SENTENCE MEANING

- We will mostly look at *meaning of sentence/utterance*
- Various relevant aspects of meaning
- Often interested in **formal** representations
- Structures composed of *symbols*
- Arrangement denotes: *objects, properties, relations*

18

DESIDERATA FOR FORMAL REPRESENTATIONS

- **Canonical form:**
 - Same meaning → same meaning *representation*
- **Inference:**
 - Representation supports computational inference
 - Generally required for MR to be useful!
- **Expressiveness:**
 - Expressive enough for all expected inputs, subject matter, ...
 - All possible NL utterances?
 - Probably not from one MR system!

20

MORE SENTENCE MEANING

- Much of relevant utterance meaning goes beyond this
- Extend to capture other aspects of meaning
- So far: **states** – objects/entities, conditions, properties
- **Events** – changes over time

Example

The ball is on the table Simple static state
The ball fell from the table Change in state

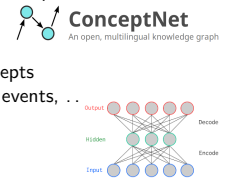
- Solution: introduce variables ranging over *events*

$$\exists e, x, y. \text{fall}(x, y, e) \wedge \text{ball}(x) \wedge \text{table}(y)$$

22

EXAMPLE MRs

- **Formal MRs:** logical expressions over entities
 $\exists e, y. \text{Having}(e) \wedge \text{Haver}(e, \text{Speaker})$
 - More in a moment
- **Network-based representations:** *semantic nets*
 - Network of semantic *relations* between *concepts*
 - E.g. WordNet, ConceptNet
- **Feature-based representations**
 - *Attribute-value* pairs for *objects* or *concepts*
 - Words, relations to other objects, facts, events, ...
- **Latent feature-based representations**
 - Vector representations
 - E.g. word embeddings
 - E.g. deep learning-based reprs: auto-encoders, sentence encoders, image features, ...



17

DESIDERATA FOR FORMAL REPRESENTATIONS

- **Verifiability:**
 - Ability to test *assertions* or *questions* against *knowledge*
- **Unambiguity:**
 - Linguistic expressions ambiguous, but MR should not be
 - Must be able to reason over clear interpretations
 - *Expression* may have many *meanings*
- **Vagueness:**
 - Some information may be left unspecified
 - Not ambiguity: single interpretation in input
 - E.g. *I want to eat Italian food* vs *I want to eat pizza*

19

TRUTH-CONDITIONAL SEMANTICS

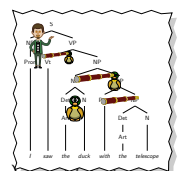
- One common form of *formal MR*
- Semantics of sentence \simeq conditions for it to be **true**
- **Conditions** defined in terms of **entities, properties, states**

The ball is on the table
 \equiv
The table is under the ball

- Typically uses **predicate logic**
 $\exists x, y. \text{on}(x, y) \wedge \text{ball}(x) \wedge \text{table}(y)$
- **First-order logic**

21

FROM TEXT TO MEANING



The ball is on the table $\mapsto \exists x, y. \text{on}(x, y) \wedge \text{ball}(x) \wedge \text{table}(y)$
depends on identifying structure of connections between words
 \rightarrow **syntactic structure** of sentence

ball is subject of *is*, ...

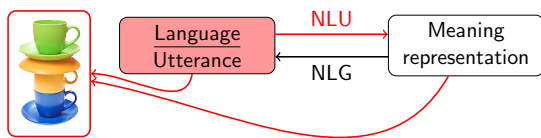
Linear input (text) \mapsto syntactic structures
Linear input (text) \mapsto (syntax) \mapsto MR

Parsing
Semantic parsing

Much more in L6 and later

24

TWO PERSPECTIVES



Meaning of utterance

Dialogue actions, sentiment/stance analysis, beliefs/knowledge in dialogue, question answering, commands, ...

State of the world

Knowledge acquisition by IE, dialogue modelling, summarization, news reporting, ...

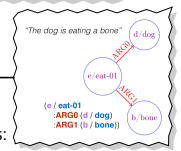
Connection is link between **linguistic input** and **real world**

25

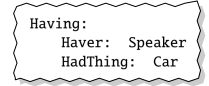
FORMAL MRs

$\exists x, y. on(x, y) \wedge ball(x) \wedge table(y)$

- First-order logic
- Abstract meaning representation (AMR)
 - Directed graph, similar to logic
 - Flexible representation, various meaning types: arguments, semantic relations, quantities, dates, ...



- Frame-based/slot-filler representations
 - Knowledge representation based on *stereotyped situations*
 - Known slots to be filled for situation
 - *Semantic role labelling* (SRL):
Text \Rightarrow Filled frame



26

NON-LOGICAL REPRESENTATIONS

Not all meaning is formal/logical

- Meaning is not discrete

$angry \stackrel{?}{=} furious$

$irritated < annoyed < angry < furious < outraged$

- Similarity
- Multi-dimensional meanings/relationships
- **Sentiment**: orientation of author toward object
- Ways to capture some other types of meaning in later lectures: e.g. word embeddings, sentence vectors

27

EXERCISE: HOW TO REPRESENT?

- **Task**: Modelling the described scene
- **Input**:

Example

In one corner of the room was a small, darkish-brown mouse. Opposite it, to my left, was a large creature, unfamiliar to me.

- Return to earlier list of relevant elements of meaning
- What **types of MR** could be used?
- (Again: don't worry about *how* to extract them)

28

MEANING REPRESENTATIONS

Summary

- Dual perspective on *meaning*:
 1. Meaning of utterance
 2. State of world
 - Connection links **language** to **real world**
- Types of MR
 - Formal: e.g. truth-conditional / FOL
 - Networks
 - Features
 - *Latent* features: e.g. word embeddings

More later on *vector representations* for many purposes and feature / network representations of words

29

REGULAR EXPRESSION AND FSAs

Coming up:

- Brief reminder of **regular expressions**
- Theory and notation for **finite-state automata**
- Introduction to **morphological analysis**
- Some uses in NLP

30

REGULAR EXPRESSIONS

Brief reminder

Pattern	Matches
/radio/	'It is the radio . Know then, O Queen
/[Rr]adio/	late that night, the Radio Man
/[sbt]ack/	Crota was already back in the fray
/[0-9]/	showed the time to be 1025 ;
/radio sets?/	powerful radio sets invented by the beast
/([Dd]it [Dd]ah)/	returning to the hidden radio set , whence ' Dah-dit-dah-dit dah-dah-dit-dah .'

31

REGULAR EXPRESSIONS

Brief reminder

Pattern	Matches
/final*/	we finally restored it,
/final+ /	we finally restored it,
/radio./	conventional radioese , I repeated
/(it ah)(-.(it ah))*/	' Dah-dit-dah-dit dah-dah-dit-dah .'

- Need more of a reminder? *Jurafsky & Martin 3, 2.1*
- Some common regex features not technically *regular*
E.g. memory

32

REGULAR LANGUAGES

- *Regular expression* defines a **regular language**
- Set of strings accepted by regex

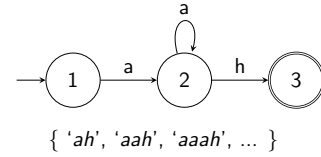
$$L(r) = \{s | \text{accept}(r, s)\}$$

- Language is *regular* iff \exists regex for it
- Regular languages describe *some* aspects of NL
- Useful tool for some NLP
- Particularly: preprocessing & early stages of analysis
- See limitations tomorrow

33

FINITE-STATE AUTOMATA

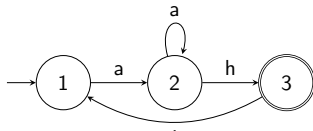
- Another string-testing formalism: **finite-state automaton (FSA)**
- Process string by transitioning between states



- Equivalent regex: `/a+h/`

34

ANOTHER FSA



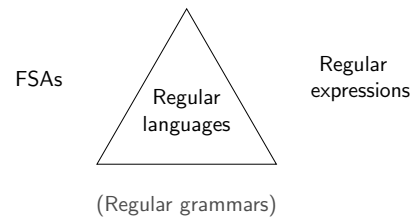
{ 'ah, ah', 'ah, aah', 'aaaaah, ah, aah', ... }

- Equivalent regex: `/a+h(, a+h)*/`

35

FSAs & REGEXES

- Acceptance by FSA \equiv acceptance by regex
- Every FSA has equivalent regex



36

FINITE STATE TRANSDUCERS

- Extend FSA to *output* something: not just YES/NO
- Each accepting edge also outputs
- Finite-state **transducer**
- Same strings/language as FSA
- Output may be:
 - translation
 - analysis
 - spelling transformation
 - ...

37

EXAMPLE TEXT

Example

In conventional radioese, I repeated the sounds to the Harvard group:

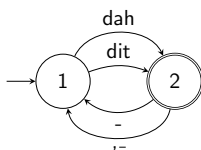
"Dah-dit-dah-dit dah-dah-dit-dah. Dah-dit-dah-dit dah-dah-dit-dah. Dah-dit-dit dit. Dah-dit-dah-dit dit-dah dah-dit dit dit dah-dah-dah dah. Dah-dit-dah-dit dit-dah dah-dit-dit-dit dah-dah-dah dah. Dah-dit-dah-dit dit-dah dah-dit-dit-dit-dah dah-dah-dah."

A look of incredulity spread over their faces. Again came the same message, and again I repeated it.

38

FSA \rightarrow FST

FSA to accept *dit-dah* sequences:

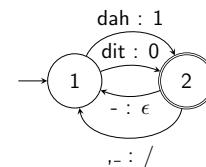


{ 'dit-dah-dah, dah-dit', ... }

`/(dit|dah)(-(dit|dah))*((, (dit|dah)(-(dit|dah))**)/`

39

FSA \rightarrow FST



- Each state: *input : output*
- Translates

dit-dah-dah, dah-dit, dit \Rightarrow *0-1-1/1-0/1*

- Common use in NLP: analysis of internal word structure \rightarrow *morphology*

40

MORPHOLOGY: SOME CONCEPTS

- **Morpheme**: smallest grammatical unit in language
- **Affix**: morpheme that occurs only together with others
- Word = **stem** [+ **affixes**]
radios → **radio** + **s**
- **Compound**: word with multiple stems
thunderstorms → **thunder** + **storm** + **s**

41

MORPHOLOGY: SOME CONCEPTS

Types of affix:

- **prefix**: **un**+help+ful
- **suffix**: taste+ful, taste+ful+ness
- **infix**: internal affix (e.g. Arabic)
- **circumfix**: prefix & suffix
E.g. German: **ge**+kauf+**t**

42

MORPHOLOGY: SOME CONCEPTS

Two types of morphology:

1. **Inflectional**: regular patterns for word classes
 - Changes *grammatical* roles
 - E.g. noun cases: *kauppa*, *kauppa+a*, *kaupa+n*, ...
2. **Derivational**: creates new words
 - Changes *meaning*
 - E.g. diminutive suffix: *tuuli* → *tuulo+nen*

43

MORPHOLOGICAL AMBIGUITY

- **Morpheme-level**:
 - Morphemes can have multiple interpretations/uses
 - *table*: noun or verb
 - *+s*: plural noun or 3rd-person singular verb
- **Structural**:
 - Words may be split in multiple ways
 - *unionised* → *union+ise+ed* / *un+ion+ise+ed*
- **Bracketing**:
 - Same split, different bracketing structures
 - *unlockable* → *(un+lock)+able* / *un+(lock+able)*

44

USES IN NLP

A few uses of morphology in NLP:

- Morphosyntactic categorization (rough POS tagging)
- Morphological features
- Stemming/lemmatization
- Generation: apply syntax, features, agreement to base forms

the loyal princes occupied the throne in his absence
det adj noun verb det noun prep det noun
adv adj prn

45

USES IN NLP

A few uses of morphology in NLP:

- Morphosyntactic categorization (rough POS tagging)
- Morphological features
- Stemming/lemmatization
- Generation: apply syntax, features, agreement to base forms

the loyal princes occupied the throne in his absence
def-sg pl past def-sg sg m-sg sg
def-pl pst-prt def-pl

45

USES IN NLP

A few uses of morphology in NLP:

- Morphosyntactic categorization (rough POS tagging)
- Morphological features
- Stemming/lemmatization
- Generation: apply syntax, features, agreement to base forms

the loyal prince occupied the throne in his absence
the loyal **prince** **occupy** the throne in his absence

45

USES IN NLP

A few uses of morphology in NLP:

- Morphosyntactic categorization (rough POS tagging)
- Morphological features
- Stemming/lemmatization
- Generation: apply syntax, features, agreement to base forms

the loyal prince occupy the throne in prn absence
+pl +pl +pl +past +sg +sg +pos +sg
+masc+sg
the loyal princes occupied the throne in his absence

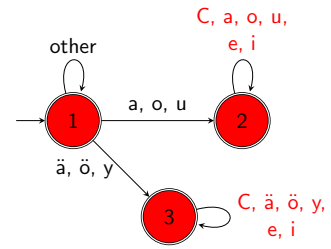
45

FSTs FOR MORPHOLOGY

- Can encode morphological **rules** as FSTs
- Example: Finnish vowel harmony
- A reminder:
 - **Back** vowels: *a, o, u*
 - **Front** vowels: *ä, ö, y*
 - **Middle** vowels: *e, i*
 - Word contain **back+middle** or **front+middle**
 - Never: **back+front**
 - Affixes come in back and front forms: e.g. *na/nä*

FSA FOR FINNISH VOWEL HARMONY

Accepts only valid combinations of front/back/middle



46

47

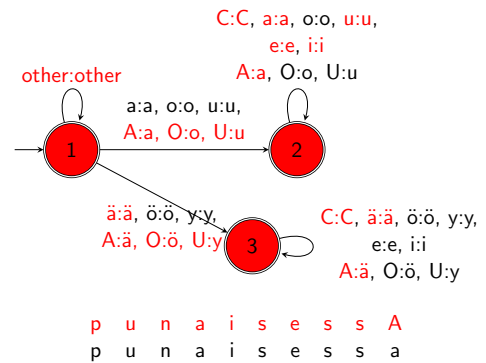
FST FOR FINNISH VOWEL HARMONY

- Affixes added by other process
- Generic form: harmony left till later (now)

A → a/ä	O → o/ö	U → u/y
---------	---------	---------

punaise+ssA → *punaisessa*
pää+ssA → *päässä*

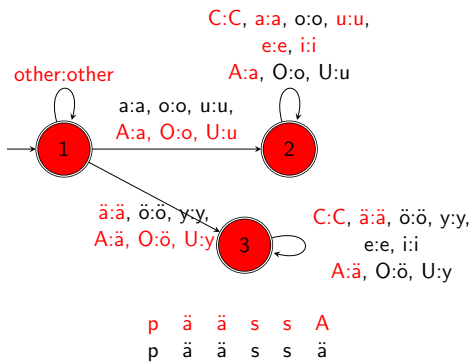
FST FOR FINNISH VOWEL HARMONY



48

49

FST FOR FINNISH VOWEL HARMONY



49

OTHER USES OF FSTs

- Morphological **generation**
- Text preprocessing, tokenization, NER, ...
- Simple dialogue systems: flow of possible questions, responses, ...
- Dialogue models: tracking dialogue state, user knowledge

More on dialogue in lecture 11

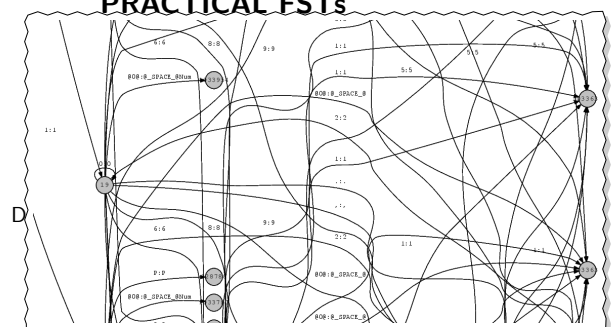
50

PRACTICAL FSTs

- Vowel harmony: very simple example, not even complete
- Real morphology much more complex
- Huge, complicated FSTs
- **Divide** into smaller components: **compose**
- Limited ambiguity: few possible analyses per word
- Real morphological analysis for *Finnish, English, Swedish, Turkish, Italian, German, ...*

HFST: <https://github.com/hfst>
[Web demo](#)

PRACTICAL FSTs



51

52

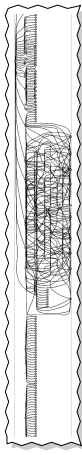
PRACTICAL FSTs

<https://github.com/hfst>

Downloadable FST for Finnish morphology analysis

~1k nodes →

Full analyser contains **2.5M** nodes!



52

FINITE-STATE METHODS

Summary

- Regular languages: regular expressions \equiv FSA
- Finite-state **transducers**
- Intro to **morphology**
- FSTs for morphology
- Other uses in NLP
- Some drawbacks
- **Next lecture:** further limitations of finite-state methods

53

READING MATERIAL

- Formal meaning representations: *J&M3 chap 16*
- More on formal representations and inference:
Book: *Representation and Inference for Natural Language*,
Blackburn & Bos, 2005
- **Regular expressions** and **FSAs**: *J&M3 2.1*
- **Morphology**: *J&M3 2.4.4 (J&M2 p79-80)*

54