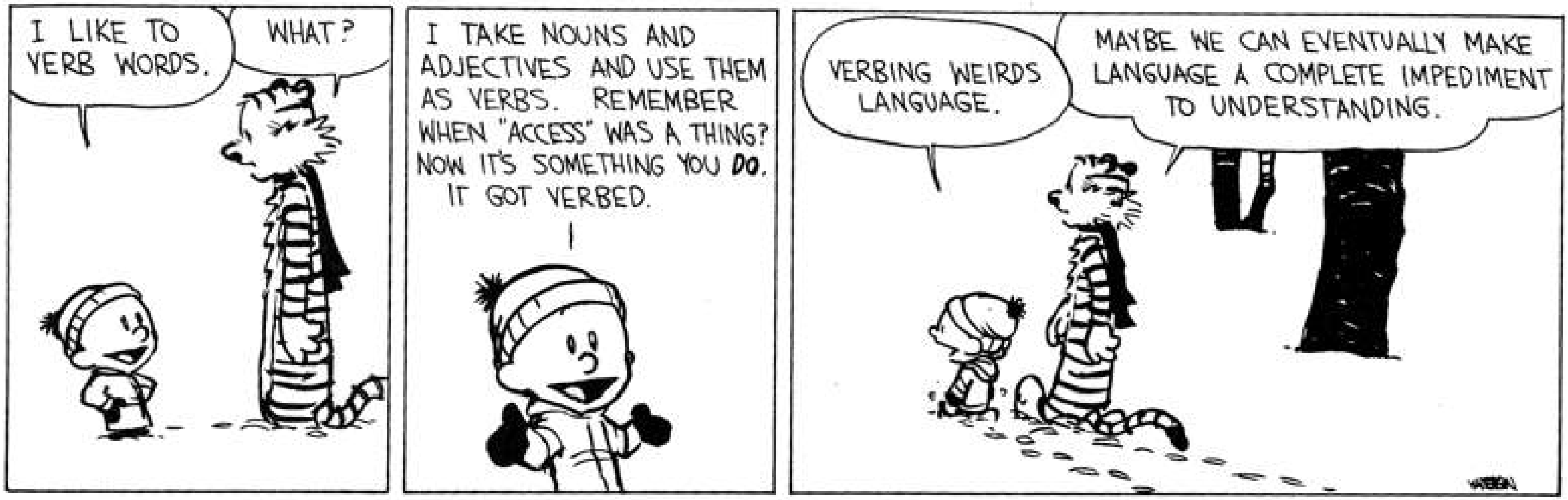
CMSC 723: Computational Linguistics I — Session #4

Part-of-Speech Tagging

Jimmy Lin
The iSchool
University of Maryland

Wednesday, September 23, 2009



Source: Calvin and Hobbs

Today's Agenda

- What are parts of speech (POS)?
- What is POS tagging?
- Methods for automatic POS tagging
  - Rule-based POS tagging
  - Transformation-based learning for POS tagging
- Along the way…
  - Evaluation
  - Supervised machine learning

Parts of Speech

- “Equivalence class” of linguistic entities
  - “Categories” or “types” of words
- Study dates back to the ancient Greeks
  - Dionysius Thrax of Alexandria (*c.* 100 BC)
  - 8 parts of speech: noun, verb, pronoun, preposition, adverb, conjunction, participle, article
  - Remarkably enduring list!

How do we define POS?

- By meaning
  - Verbs are actions
  - Adjectives are properties
  - Nouns are things

Unreliable! Think back to the comic!

- By the syntactic environment
  - What occurs nearby?
  - What does it act as?
- By what morphological processes affect it
  - What affixes does it take?
- Combination of the above

Parts of Speech

- Open class
  - Impossible to completely enumerate
  - New words continuously being invented, borrowed, etc.
- Closed class
  - Closed, fixed membership
  - Reasonably easy to enumerate
  - Generally, short function words that “structure” sentences

Open Class POS

- Four major open classes in English
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs
- All languages have nouns and verbs... but may not have the other two

Nouns

- Open class
  - New inventions all the time: muggle, webinar, ...
- Semantics:
  - Generally, words for people, places, things
  - But not always (bandwidth, energy, ...)
- Syntactic environment:
  - Occurring with determiners
  - Pluralizable, possessivizable
- Other characteristics:
  - Mass vs. count nouns

Verbs

- Open class
  - New inventions all the time: google, tweet, ...
- Semantics:
  - Generally, denote actions, processes, etc.
- Syntactic environment:
  - Intransitive, transitive, ditransitive
  - Alternations
- Other characteristics:
  - Main vs. auxiliary verbs
  - Gerunds (verbs behaving like nouns)
  - Participles (verbs behaving like adjectives)

Adjectives and Adverbs

- Adjectives
  - Generally modify nouns, e.g., *tall* girl
- Adverbs
  - A semantic and formal potpourri…
  - Sometimes modify verbs, e.g., sang *beautifully*
  - Sometimes modify adjectives, e.g., *extremely* hot

Closed Class POS

- Prepositions
  - In English, occurring before noun phrases
  - Specifying some type of relation (spatial, temporal, …)
  - Examples: *on* the shelf, *before* noon
- Particles
  - Resembles a preposition, but used with a verb ("phrasal verbs")
  - Examples: find *out*, turn *over*, go *on*

Particle vs. Prepositions

He came *by* the office in a hurry	(by = preposition)
He came *by* his fortune honestly	(by = particle)
We ran *up* the phone bill	(up = particle)
We ran *up* the small hill	(up = preposition)
He lived *down* the block	(down = preposition)
He never lived *down* the nicknames	(down = particle)

More Closed Class POS

- Determiners
  - Establish reference for a noun
  - Examples: *a*, *an*, *the* (articles), *that*, *this*, *many*, *such*, …
- Pronouns
  - Refer to person or entities: *he*, *she*, *it*
  - Possessive pronouns: *his*, *her*, *its*
  - Wh-pronouns: *what*, *who*

Closed Class POS: Conjunctions

- Coordinating conjunctions
  - Join two elements of “equal status”
  - Examples: cats *and* dogs, salad *or* soup
- Subordinating conjunctions
  - Join two elements of “unequal status”
  - Examples: We’ll leave *after* you finish eating. *While* I was waiting in line, I saw my friend.
  - Complementizers are a special case: I think *that* you should finish your assignment

Lest you think it's an Anglo-centric world,
It's time to visit ......

The (Linguistic)
Twilight Zone

Digression

The (Linguistic)Twilight Zone

Perhaps, not so strange…

Turkish

uygarlaştıramadıklarımızdanmışsınızcasına →
uygar+laş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına
behaving as if you are among those whom we could not cause to become civilized

Chinese

No verb/adjective distinction!
漂亮: beautiful/to be beautiful

Digression

The (Linguistic)Twilight Zone

Tzeltal (Mayan language spoken in Chiapas)

Only 3000 root forms in the vocabulary

The verb ‘EAT’ has **eight** variations:

- General : TUN
- Bananas and soft stuff : LO’
- Beans and crunchy stuff : K’UX
- Tortillas and bread : WE’
- Meat and Chilies : TI’
- Sugarcane : TZ’U
- Liquids : UCH’

Digression

The (Linguistic)Twilight Zone

Riau Indonesian/Malay

No Articles

No Tense Marking

3rd person pronouns neutral to both gender and number

No features distinguishing verbs from nouns

Digression

The (Linguistic)Twilight Zone

Riau Indonesian/Malay

Ayam (chicken) *Makan* (eat)

The chicken is eating
The chicken ate
The chicken will eat
The chicken is being eaten
Where the chicken is eating
How the chicken is eating
Somebody is eating the chicken
The chicken that is eating

Back to regularly scheduled programming…

POS Tagging: What's the task?

- Process of assigning part-of-speech tags to words
- But what tags are we going to assign?
  - Coarse grained: noun, verb, adjective, adverb, …
  - Fine grained: {proper, common} noun
  - Even finer-grained: {proper, common} noun $\pm$ animate

What's the tradeoff?

- Important issues to remember
  - Choice of tags encodes certain distinctions/non-distinctions
  - Tagsets will differ across languages!
- For English, Penn Treebank is the most common tagset

Penn Treebank Tagset: 45 Tags

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	“to”	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential ‘there’	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	“	left quote	‘ or “
POS	possessive ending	*’s*	”	right quote	’ or ”
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one’s*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Penn Treebank Tagset: Choices

- Example:
  - The/DT grand/JJ jury/NN commmented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- Distinctions and non-distinctions
  - Prepositions and subordinating conjunctions are tagged “IN” (“Although/IN I/PRP..”)
  - Except the preposition/complementizer “to” is tagged “TO”

Don’t think this is correct? Doesn’t make sense?

Often, must suspend linguistic intuition and defer to the annotation guidelines!

Why do POS tagging?

- One of the most basic NLP tasks
  - Nicely illustrates principles of statistical NLP
- Useful for higher-level analysis
  - Needed for syntactic analysis
  - Needed for semantic analysis
- Sample applications that require POS tagging
  - Machine translation
  - Information extraction
  - Lots more…

Why is it hard?

- Not only a lexical problem
  - Remember ambiguity?
- Better modeled as sequence labeling problem
  - Need to take into account context!

Try your hand at tagging...

- The back door
- On my back
- Win the voters back
- Promised to back the bill

Try your hand at tagging...

- I thought that you...
- That day was nice
- You can go that far

Why is it hard?*

		87-tag Original Brown		45-tag Treebank Brown	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*'s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

Part-of-Speech Tagging

- How do you do it automatically?
- How well does it work?

This first

It's all about the ~~benjamins~~ evaluation

Evolution of the Evaluation

- Evaluation by **argument**
- Evaluation by **inspection** of examples
- Evaluation by **demonstration**
- Evaluation by **improvised** demonstration
- Evaluation on **data** using a figure of merit
- Evaluation on **test data**
- Evaluation on **common** test data
- Evaluation on common, **unseen** test data

Evaluation Metric

- Binary condition (correct/incorrect):
  - Accuracy
- Set-based metrics (illustrated with document retrieval):

	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

Collection size = A+B+C+D
Relevant = A+C
Retrieved = A+B

  - Precision = A / (A+B)
  - Recall = A / (A+C)
  - Miss = C / (A+C)
  - False alarm (fallout) = B / (B+D)
  - F-measure: $F = \frac{(\beta^2+1)PR}{\beta^2 P + R}$

Components of a Proper Evaluation

- Figures(s) of merit
- Baseline
- Upper bound
- Tests of statistical significance

Part-of-Speech Tagging

- How do you do it automatically? **Now this**
- How well does it work?

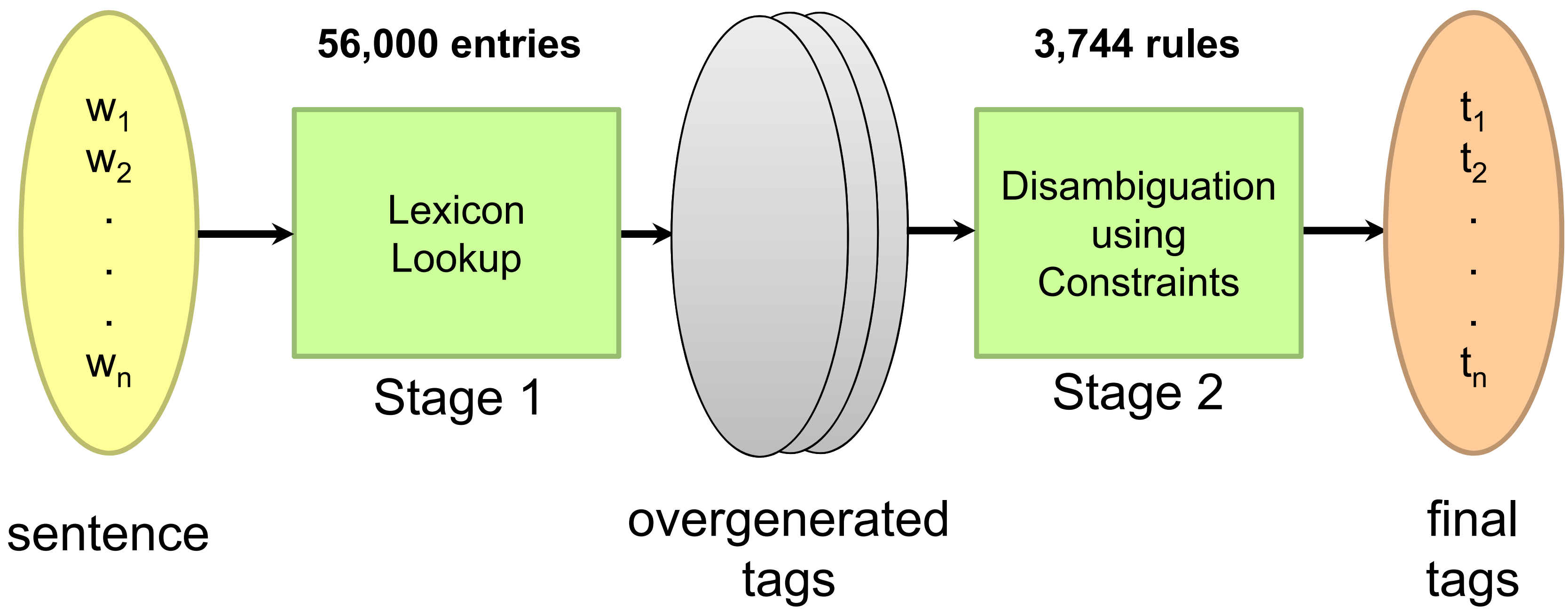
Automatic POS Tagging

- Rule-based POS tagging (now)
- Transformation-based learning for POS tagging (later)
- Hidden Markov Models (next week)
- Maximum Entropy Models (CMSC 773)
- Conditional Random Fields (CMSC 773)

Rule-Based POS Tagging

- Dates back to the 1960's
- Combination of lexicon + hand crafted rules
  - Example: EngCG (English Constraint Grammar)

EngCG Architecture



EngCG: Sample Lexical Entries

Word	POS	Additional POS features
smaller	ADJ	COMPARATIVE
fast	ADV	SUPERLATIVE
that	DET	CENTRAL DEMONSTRATIVE SG
all	DET	PREDETERMINER SG/PL QUANTIFIER
dog's	N	GENITIVE SG
furniture	N	NOMINATIVE SG NOINDEFDETERMINER
one-third	NUM	SG
she	PRON	PERSONAL FEMININE NOMINATIVE SG3
show	V	PRESENT -SG3 VFIN
show	N	NOMINATIVE SG
shown	PCP2	SVOO SVO SV
occurred	PCP2	SV
occurred	V	PAST VFIN SV

EngCG: Constraint Rule Application

Example Sentence: *Newman had originally practiced that ...*

Newman	**NEWMAN N NOM SG PROPER**
had	**HAVE <SVO> V PAST VFIN**
	HAVE <SVO> PCP2
originally	**ORIGINAL ADV**
practiced	PRACTICE <SVO> <SV> V PAST VFIN
	PRACTICE <SVO> <SV> PCP2
that	~~ADV~~
	PRON DEM SG
	DET CENTRAL DEM SG
	CS

overgenerated tags

```
ADVERBIAL-THAT Rule
Given input: that
if
    (+1 A/ADV/QUANT);
    (+2 SENT-LIM);
    (NOT -1 SVOC/A);
then eliminate non-ADV tags
else eliminate ADV tag
```

disambiguation constraint

I thought that you... (subordinating conjunction)
That day was nice. (determiner)
You can go that far. (adverb)

EngCG: Evaluation

- Accuracy ~96%*
- A lot of effort to write the rules and create the lexicon
  - Try debugging interaction between thousands of rules!
  - Recall discussion from the first lecture?
- Assume we had a corpus *annotated* with POS tags
  - Can we *learn* POS tagging automatically?

Supervised Machine Learning

- Start with annotated corpus
  - Desired input/output behavior
- Training phase:
  - Represent the training data in some manner
  - Apply learning algorithm to produce a system (tagger)
- Testing phase:
  - Apply system to unseen test data
  - Evaluate output

Three Laws of Machine Learning

- Thou shalt not mingle training data with test data
- Thou shalt not mingle training data with test data
- Thou shalt not mingle training data with test data

But what do you do if you need more test data?

Three Pillars of Statistical NLP

- Corpora (training data)
- Representations (features)
- Learning approach (models and algorithms)

Automatic POS Tagging

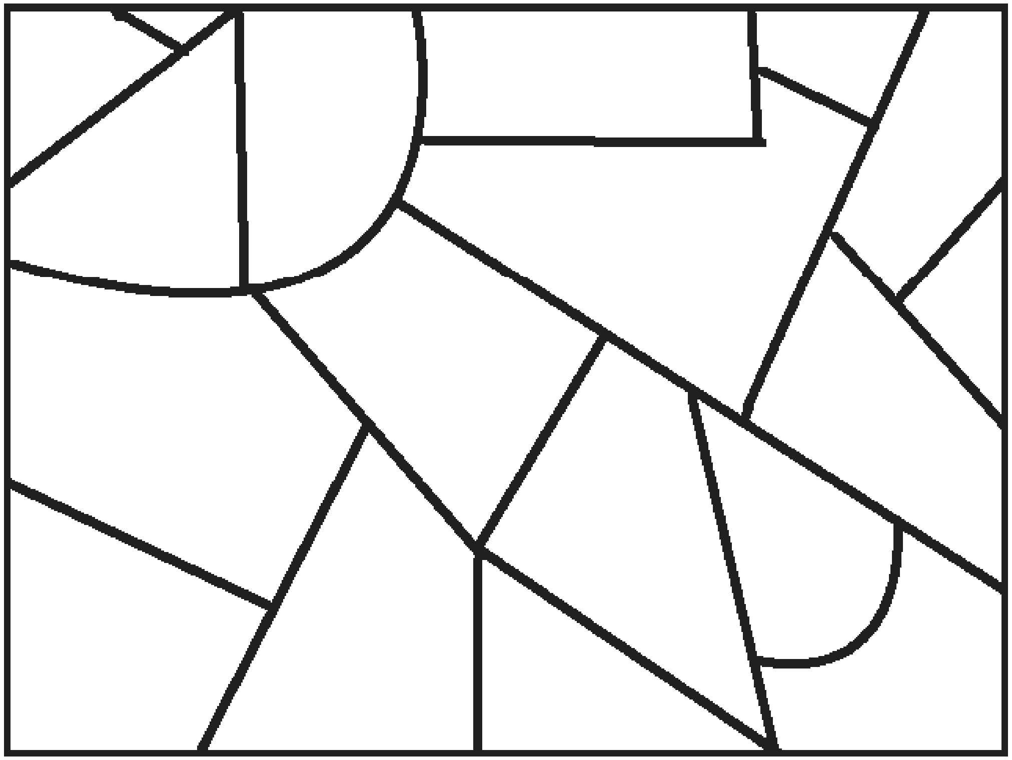
- Rule-based POS tagging (before)
- Transformation-based learning for POS tagging (now)
- Hidden Markov Models (next week)
- Maximum Entropy Models (CMSC 773)
- Conditional Random Fields (CMSC 773)

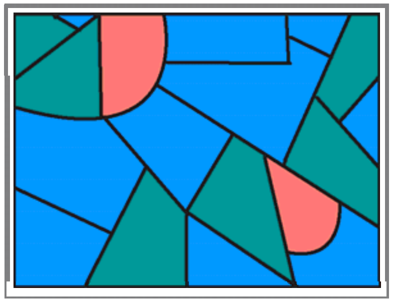
**The problem isn't with rules per se…
but with manually writing rules!**

Learn to automatically paint the next Cubist masterpiece

TBL: Training

TBL: Training



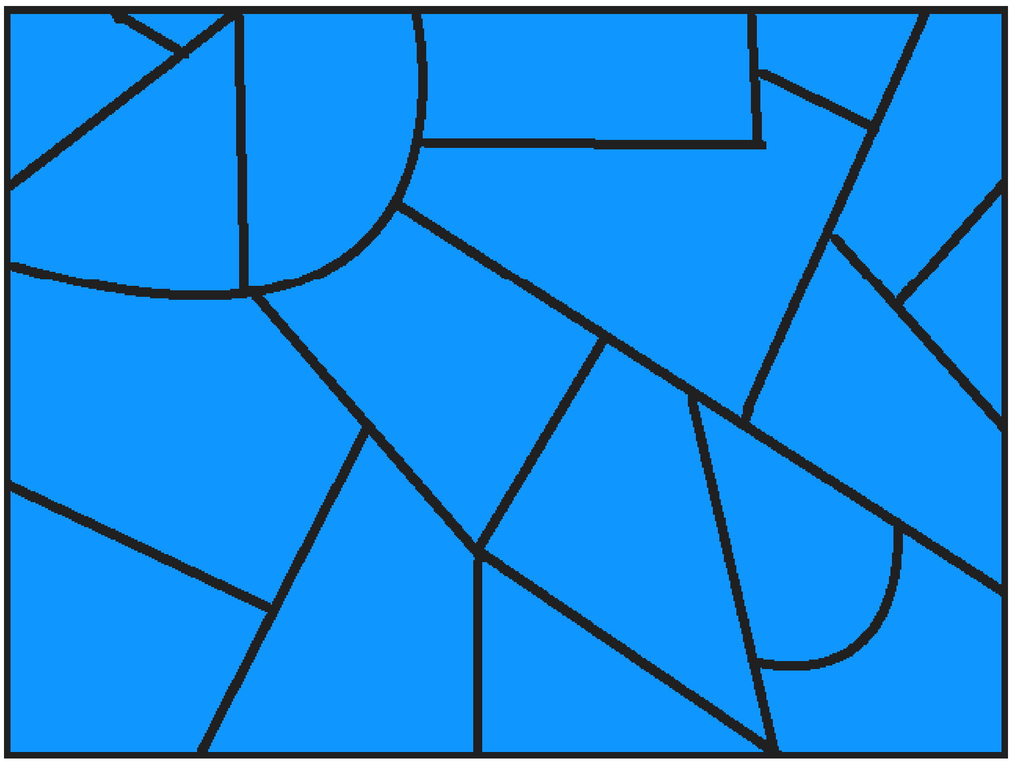


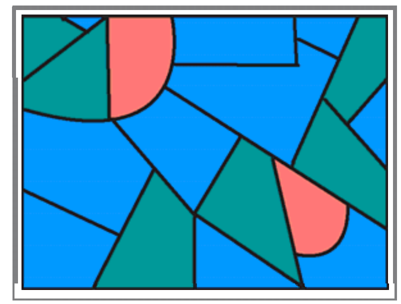
Error: 100%

Most common: **BLUE**

Initial Step: Apply Broadest Transformation

TBL: Training



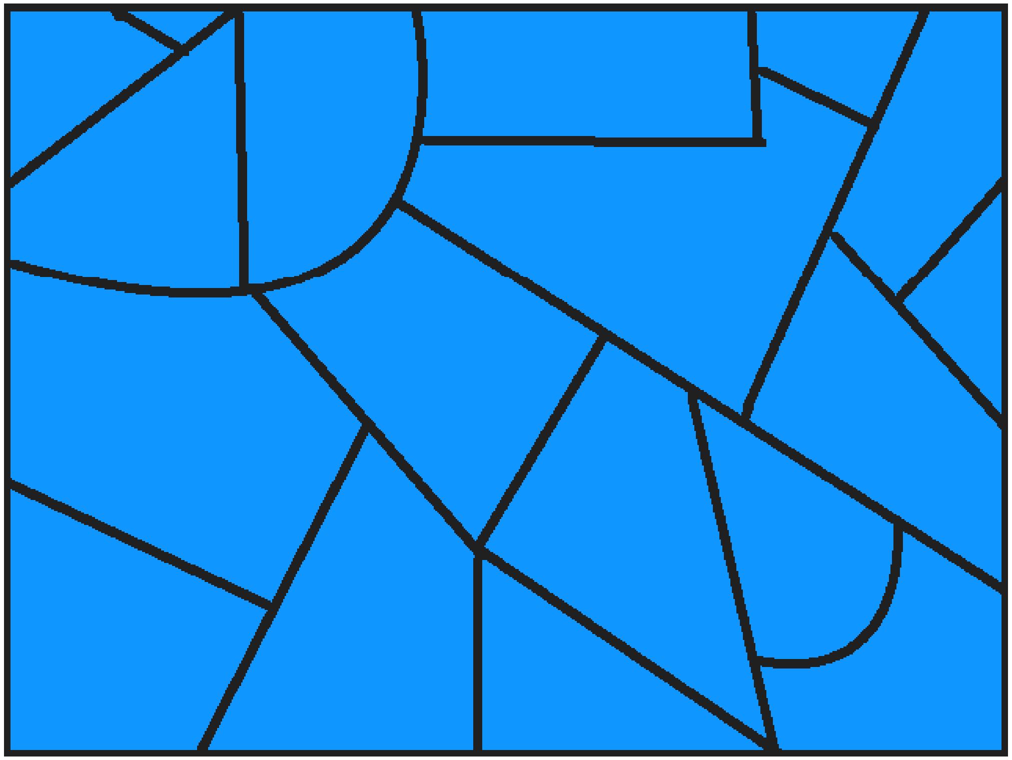


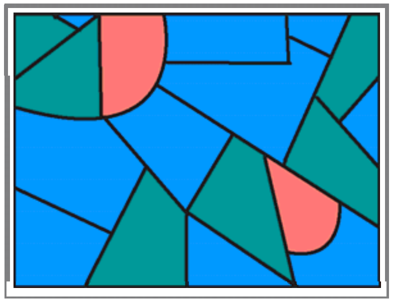
Error: 44%

`change` **B** `to` **G** `if touching` ▲

Step 2: Find transformation that decreases error most

TBL: Training



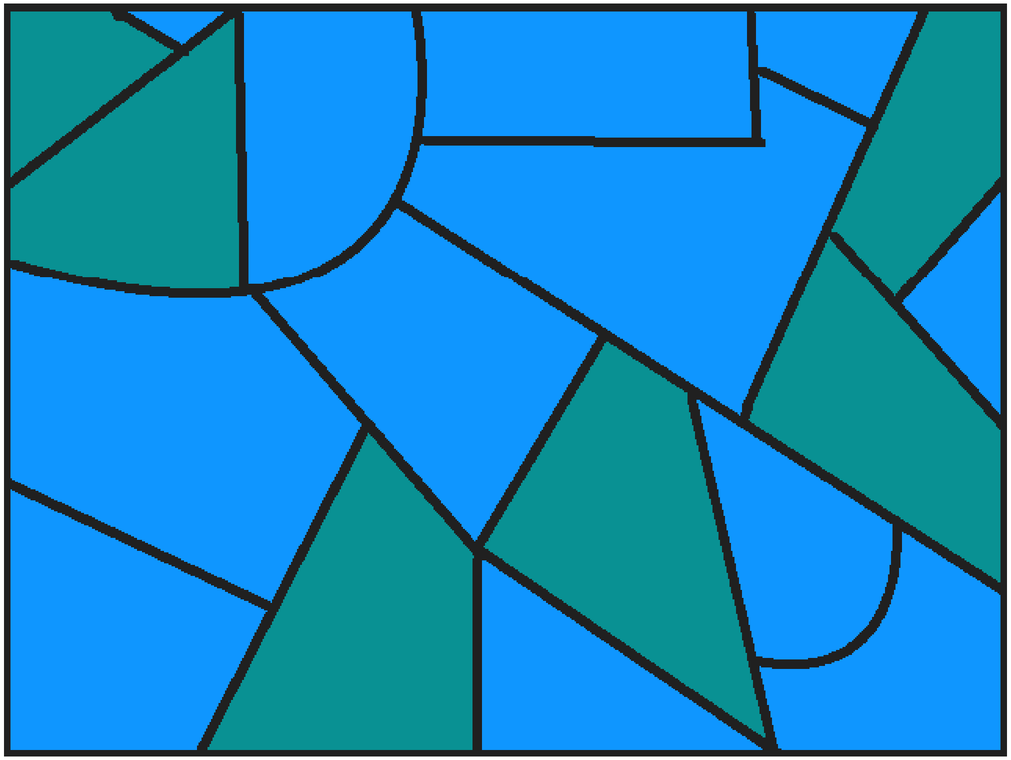


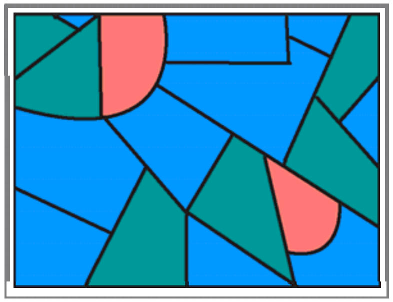
Error: 44%

change **B** to **G** if touching ▲

Step 3: Apply this transformation

TBL: Training



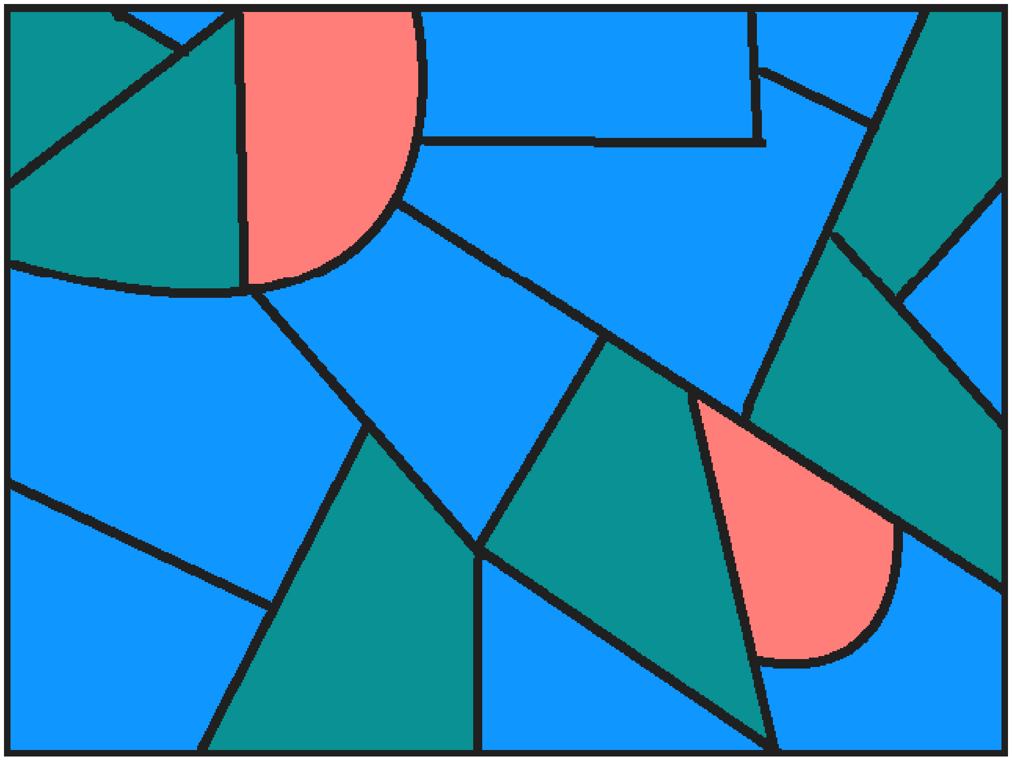


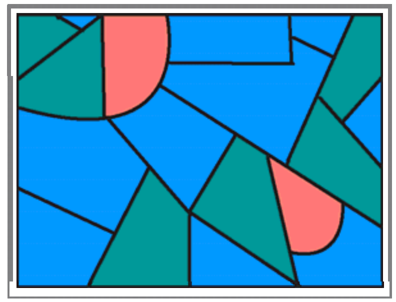
Error: 11%

change B to R if shape is

Repeat Steps 2 and 3 until “no improvement”

TBL: Training





Error: 0%

Finished !

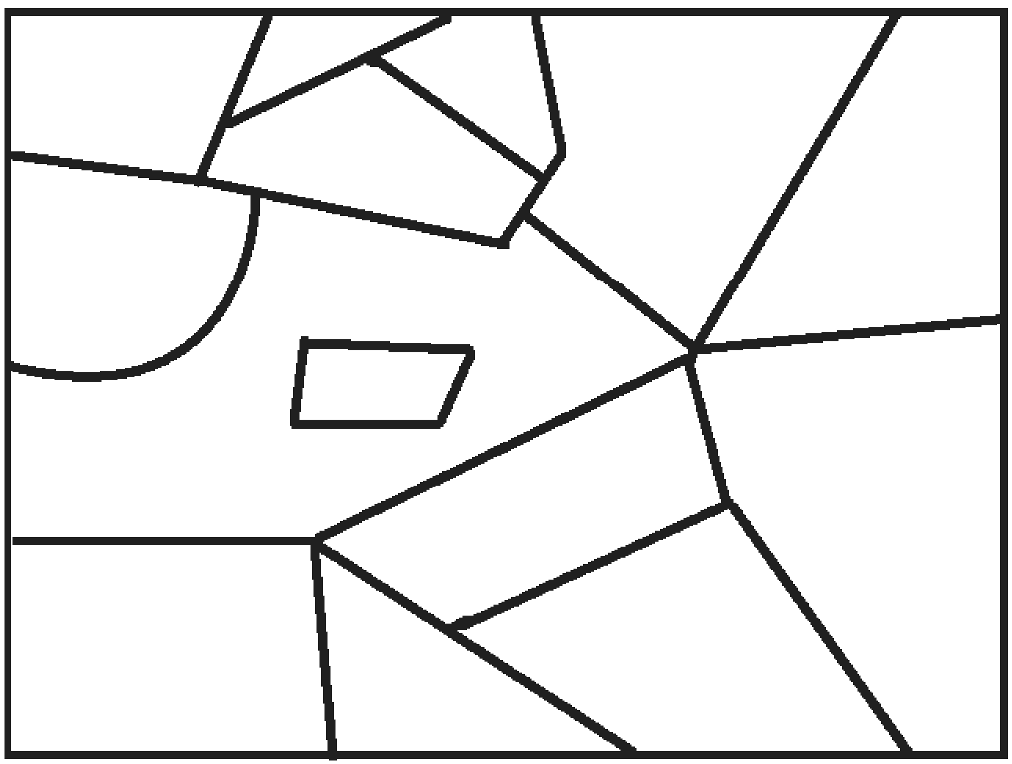
TBL: Training

- What was the point? We already had the right answer!
- Training gave us ordered list of transformation rules
- Now apply to any empty canvas!

Picasso in a box!

TBL: Testing

TBL: Testing



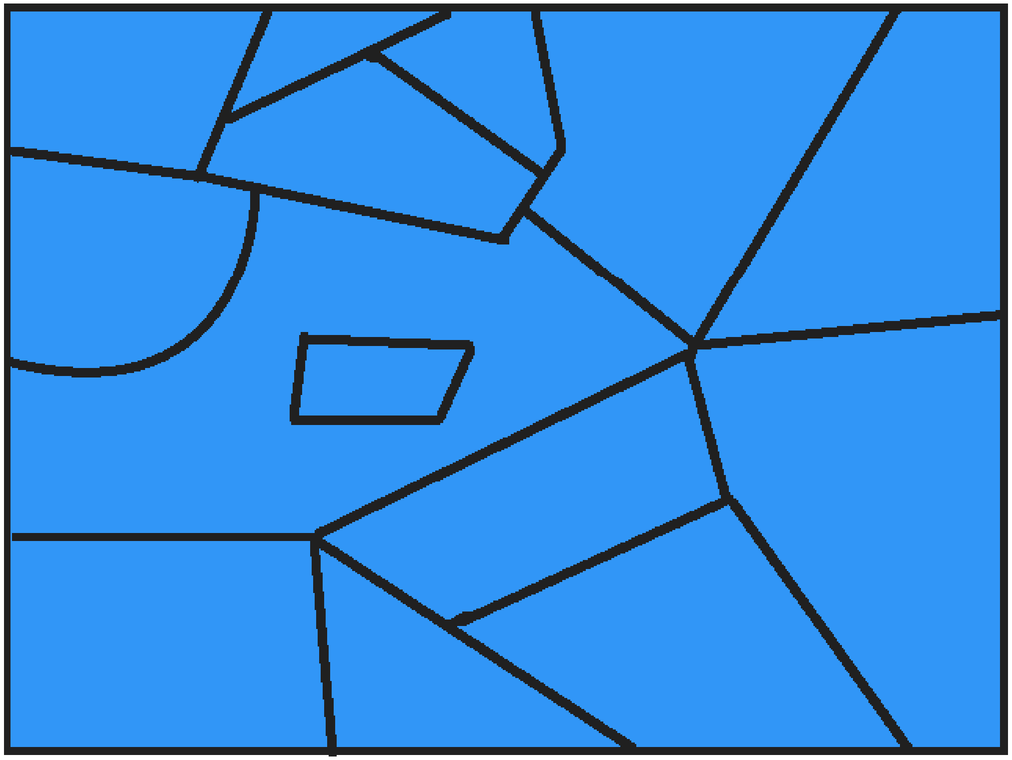
Ordered transformations:

Initial: Make all **B**

change **B** to **G** if touching ▲

change **B** to **R** if shape is

TBL: Testing



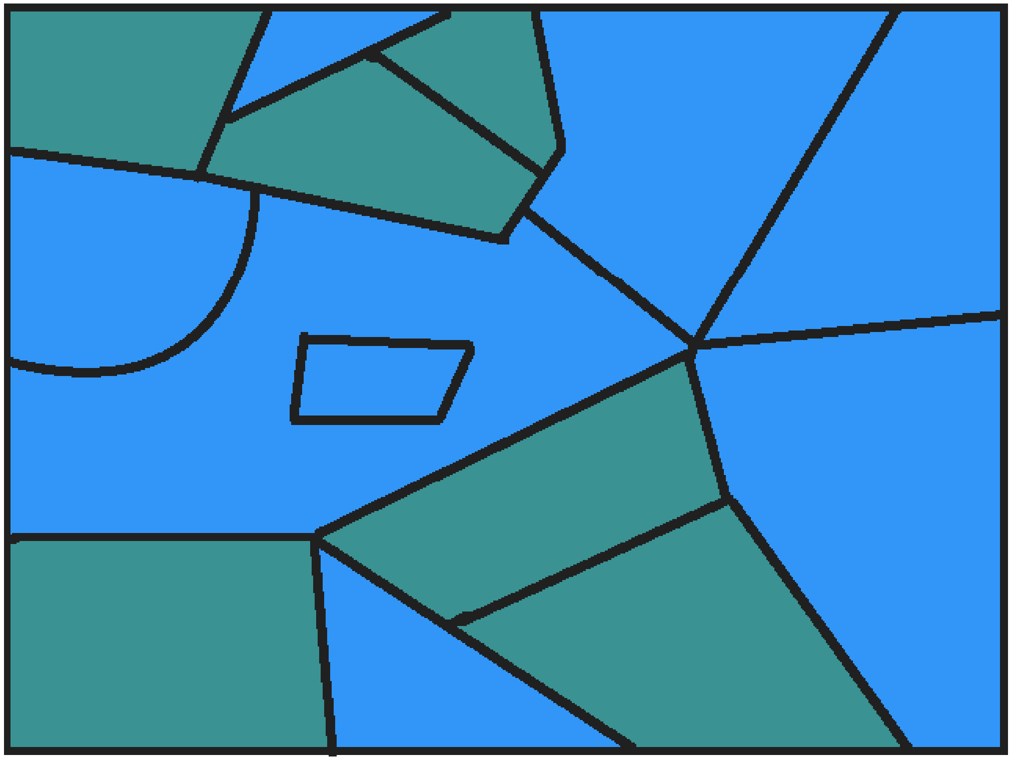
Ordered transformations:

Initial: Make all **B**

change **B** to **G** if touching

change **B** to **R** if shape is

TBL: Testing



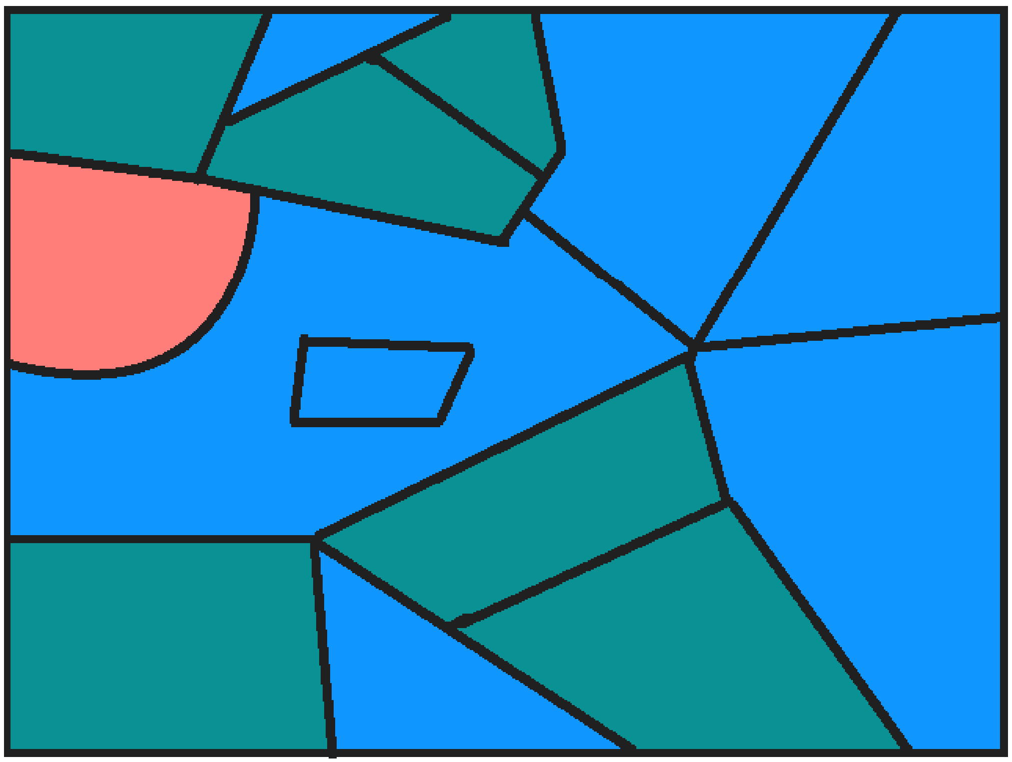
Ordered transformations:

Initial: Make all **B**

change **B** to **G** if touching

change **B** to **R** if shape is

TBL: Testing



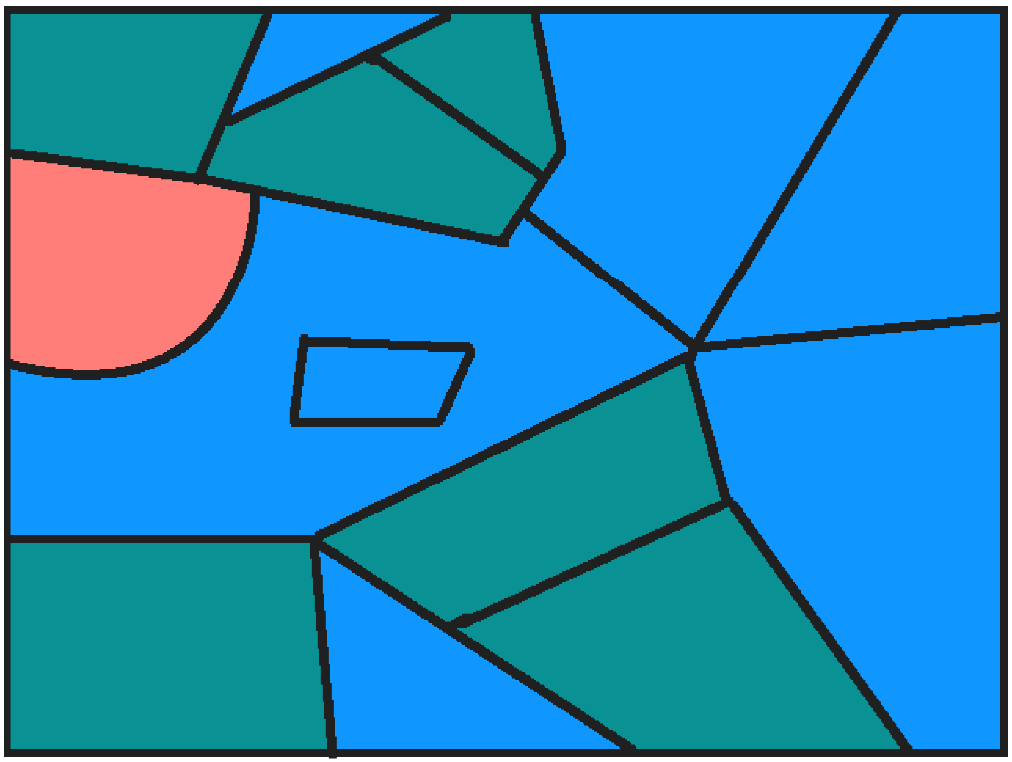
Ordered transformations:

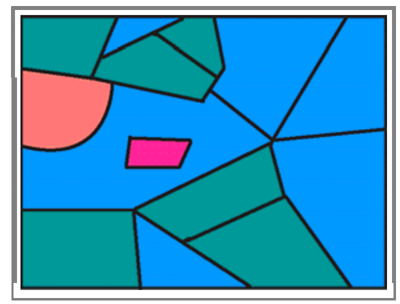
Initial: Make all **B**

change **B** to **G** if touching ▲

change **B** to **R** if shape is

TBL: Testing





Accuracy: 93%

TBL Painting Algorithm

```
function TBL-Paint
(given: empty canvas with goal painting)

begin

  apply initial transformation to canvas

  repeat

    try all color transformation rules

    find transformation rule yielding most improvements

    apply color transformation rule to canvas

  until improvement below some threshold

end
```

TBL Painting Algorithm

```
function TBL-Paint
(given: empty ca

begin

  apply initial

  repeat

    try all colo

    find transfo                                improvements

    apply color                                 vas

  until improvem

end
```

Now, substitute:

‘*tag*’ for ‘*color*’
‘*corpus*’ for ‘*canvas*’
‘*untagged*’ for ‘*empty*’
‘*tagging*’ for ‘*painting*’

TBL Painting Algorithm

```
function TBL-Paint
(given: empty canvas with goal painting)
begin
  apply initial transformation to canvas
  repeat
    try all color transformation rules
    find transformation rule yielding most improvements
    apply color transformation rule to canvas
  until improvement below some threshold
end
```

Impossible!

TBL Templates

```
Change tag t1 to tag t2 when:
   w-1 (w+1) is tagged t3
   w-2 (w+2) is tagged t3
   w-1 is tagged t3 and w+1 is tagged t4
   w-1 is tagged t3 and w+2 is tagged t4
```

Non-Lexicalized

```
Change tag t1 to tag t2 when:
   w-1 (w+1) is foo
   w-2 (w+2) is bar
   w is foo and w-1 is bar
   w is foo, w-2 is bar and w+1 is baz
```

Lexicalized

Only try instances of these (and their combinations)

TBL Example Rules

He/PRP is/VBZ as/IN tall/JJ as/IN her/PRP$

`Change from` **`IN`** `to` **`RB`** `if w+2 is` *`as`*

He/PRP is/VBZ as/RB tall/JJ as/IN her/PRP$

He/PRP is/VBZ expected/VBN to/TO race/NN today/NN

`Change from` **`NN`** `to` **`VB`** `if w-1 is tagged as` **`TO`**

He/PRP is/VBZ expected/VBN to/TO race/VB today/NN

TBL POS Tagging

- Rule-based, but data-driven
  - No manual knowledge engineering!
- Training on 600k words, testing on known words only
  - Lexicalized rules: learned 447 rules, 97.2% accuracy
  - Early rules do most of the work: 100 → 96.8%, 200 → 97.0%
  - Non-lexicalized rules: learned 378 rules, 97.0% accuracy
  - Little difference… why?
- How good is it?
  - Baseline: 93-94%
  - Upper bound: 96-97%

Source: Brill (Computational Linguistics, 1995)

Three Pillars of Statistical NLP

- Corpora (training data)
- Representations (features)
- Learning approach (models and algorithms)

In case you missed it…

Uh… what about this assumption?

- Assume we had a corpus *annotated* with POS tags
  - Can we *learn* POS tagging automatically?

Yes, as we've just shown…

knowledge engineering vs. manual annotation

Penn Treebank Tagset

- Why does everyone use it?
- What's the problem?
- How do we get around it?

Turkish Morphology

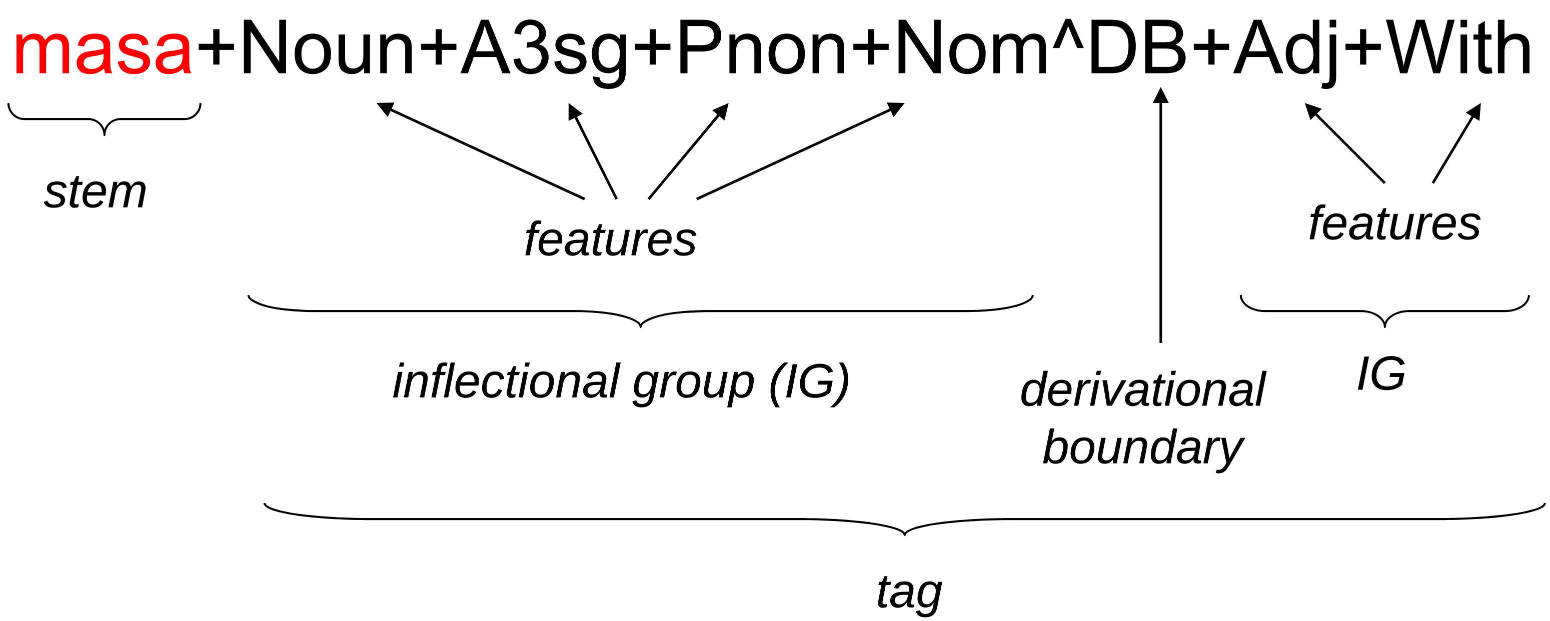
- Remember agglutinative languages?
  - uygarlaştıramadıklarımızdanmışsınızcasına → uygar+laş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına
  - *behaving as if you are among those whom we could not cause to become civilized*
- How bad does it get?
  - uyu – sleep
  - uyut – make X sleep
  - uyuttur – have Y make X sleep
  - uyutturt – have Z have Y make X sleep
  - uyutturttur – have W have Z have Y make X sleep
  - uyutturtturt – have Q have W have Z …
  - …

Source: Yuret and Türe (HLT/NAACL 2006)

Turkish Morphological Analyzer

- Example: masalı
  - masal+Noun+A3sg+Pnon+Acc (= the story)
  - masal+Noun+A3sg+P3sg+Nom (= his story)
  - masa+Noun+A3sg+Pnon+Nom^DB+Adj+With (= with tables)
- Disambiguation in context:
  - Uzun masalı anlat (Tell the long story)
  - Uzun masalı bitti (His long story ended)
  - Uzun masalı oda (Room with long table)

Morphology Annotation Scheme



- How rich is Turkish morphology?
  - 126 unique features
  - 9129 unique IGs
  - infinite unique tags
  - 11084 distinct tags observed in 1M word training corpus

How to tackle the problem...

- Key idea: build separate decision lists for each feature
- Sample rules for +Det:

R1	If	(W = çok) and (R1 = +DA)	
	Then	W has +Det	
R2	If	(L1 = pek)	
	Then	W has +Det	
R3	If	(W = +AzI)	
	Then	W does not have +Det	
R4	If	(W = çok)	
	Then	W does not have +Det	
R5	If	TRUE	
	Then	W has +Det	

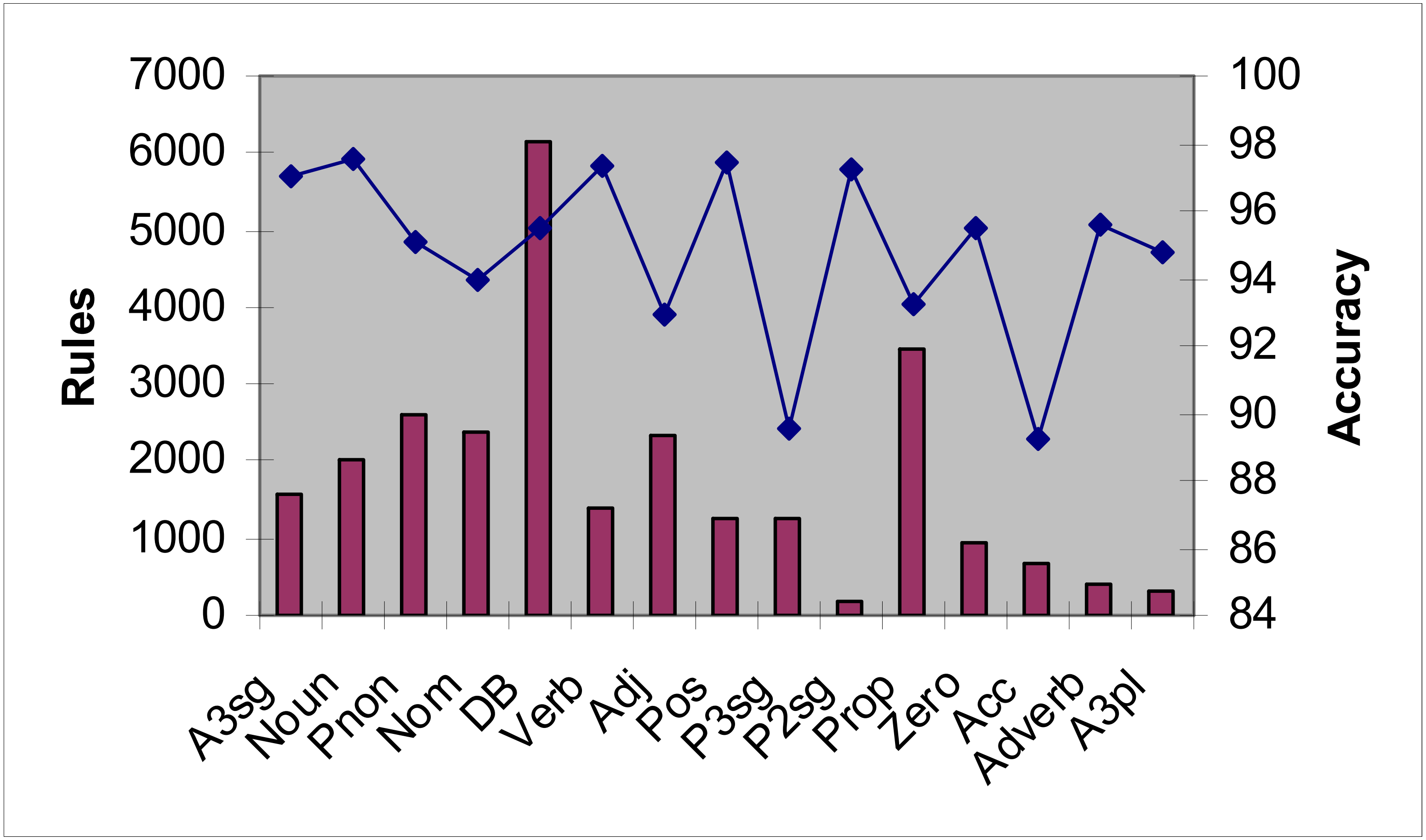
- “pek çok alanda” (R1)
- “pek çok insan” (R2)
- “insan çok daha” (R4)

Learning Decision Lists

- Start with tagged collection
  - 1 million words in the news genre
- Apply greedy-prepend algorithm
  - Rule templates based on words, suffixes, character classes within a five word window

```
GPA(data)
dlist = NIL
default-class = Most-Common-Class(data)
rule = [If TRUE Then default-class]
while Gain(rule, dlist, data) > 0
    do dlist = prepend(rule, dlist)
       rule = Max-Gain-Rule(dlist, data)
return dlist
```

Results



Overall accuracy: ~96%!

What we covered today…

- What are parts of speech (POS)?
- What is POS tagging?
- Methods for automatic POS tagging
  - Rule-based POS tagging
  - Transformation-based learning for POS tagging
- Along the way…
  - Evaluation
  - Supervised machine learning