Topic Segmentation

Regina Barzilay

February 8, 2004

What is Segmentation?

Segmentation: determining the positions at which topics change in a stream of text or speech.

SEGMENT 1: OKAY
tsk There's a farmer,
he looks like ay uh Chicano American,
he is picking pears.
A-nd u-m he's just picking them,
he comes off the ladder,
a-nd he- u-h puts his pears into the basket.
SEGMENT 2: U-h a number of people are going by,
and one of them is um I don't know,
I can't remember the first . . . the first person that goes by

Motivation

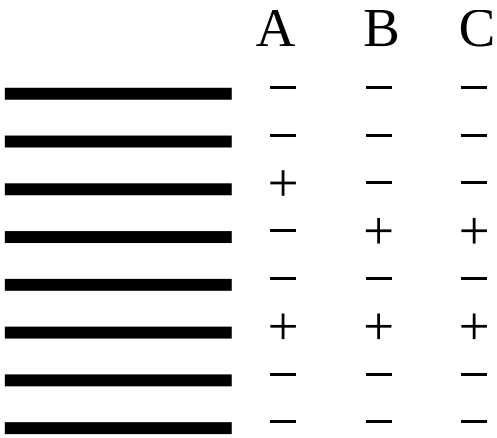
- Information Retrieval
- Summarization
- Question-Answering
- Word-sense disambiguation and anaphora resolution

Today's Topics

- Human Agreement on Segmentation and Evaluation
- Segmentation Algorithms:
  - Features: word distribution, cue words, speaker, change,. . .
  - Methods: classification, clustering, HMMs, . . .
- Segmentation for different genres: text, meetings, broadcasts,

Segmentation: Agreement

Percent agreement — ratio between observed agreements and possible agreements



$$\frac{22}{8 * 3} = 91\%$$

Results on Agreement

Grosz&Hirschbergberg'92	newspaper text	74-95%
Hearst'93	expository text	80%
Passanneau&Litman'93	monologues	82-92%

Cochran's Test

Estimate the null hypothesis that the number of subjects assigning a boundary at any position is randomly distributed

Evaluation Measures

	Boundary	Non-boundary
Alg. Boundary	a	b
Alg. Non-boundary	c	d

Recall $\frac{a}{a+c}$

Precision $\frac{a}{a+b}$

Error $\frac{b+c}{a+b+c+d}$

Simple Algorithm

Passanneau&Litman'93

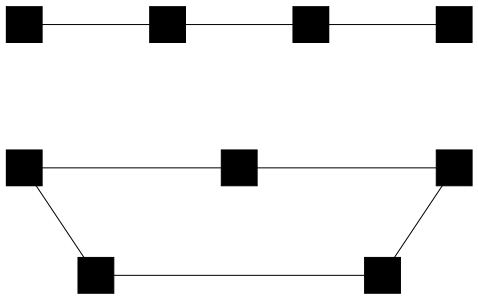
	Recall	Precision	Error
Cue	72%	15%	50%
Pause	92%	18%	49%
Humans	74%	55%	11%

Text Segmentation

Hearst'94

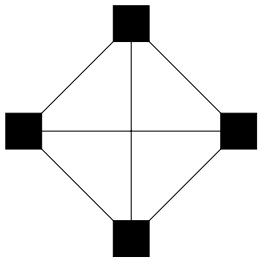
- Goal: divide text into coherent segments
- Main Idea: change in lexical connectivity patterns signals topic change
  - Linguistic Theory: Text Cohesion

Skorochodko's Text Types

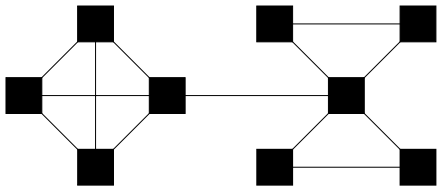


Chained

Ringed



Monolith



Piecewise

Flow model of discourse

Chafe'76:

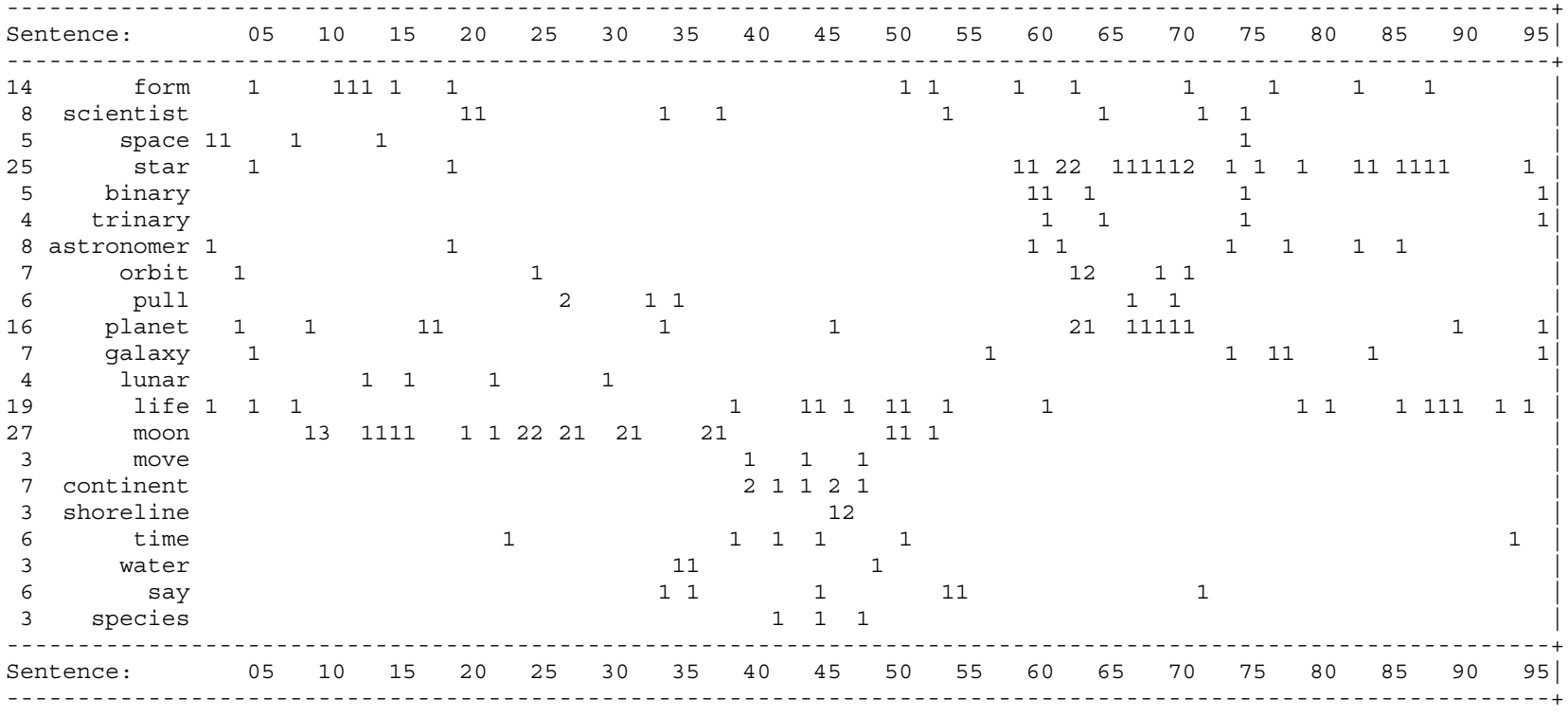
> "Our data ... suggest that as a speaker moves from focus to focus (or from thought to thought) there are certain points at which they may be a more or less radical change in space, time, character configuration, event structure, or even world ... At points where all these change in a maximal way, an episode boundary is strongly present."

Example

Stargazers Text(from Hearst, 1994)

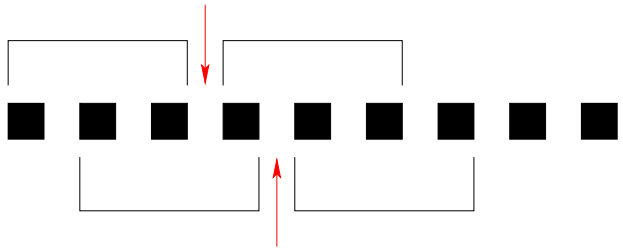
- Intro - the search for life in space
- The moon's chemical composition
- How early proximity of the moon shaped it
- How the moon helped the life evolve on earth
- Improbability of the earth-moon system

Example


Sentence: 05 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
14 form 1 111 1 1 1 1 1 1 1 1
8 scientist 11 1 1 1 1 1 1
5 space 11 1 1 1
25 star 1 1 1 11 22 111112 1 1 1 11 1111 1
5 binary 11 1 1 1
4 trinary 1 1 1 1
8 astronomer 1 1 1 1 1 1 1 1
7 orbit 1 1 12 1 1
6 pull 2 1 1 1 1
16 planet 1 1 11 1 1 21 11111 1 1
7 galaxy 1 1 1 11 1 1
4 lunar 1 1 1 1 1
19 life 1 1 1 1 11 1 11 1 1 1 1 1 111 1 1
27 moon 13 1111 1 1 22 21 21 21 11 1
3 move 1 1 1
7 continent 2 1 1 2 1
3 shoreline 12
6 time 1 1 1 1 1 1
3 water 11 1 1
6 say 1 1 1 11 1
3 species 1 1 1
Sentence: 05 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

Segmentation Algorithm

- Preprocessing and Initial segmentation
- Similarity Computation
- Boundary Detection



Preprocessing and Initial Segmentation

- Tokenization
- Morphological analysis
- Token-sequence division

Similarity Computation: Representation

Vector-Space Representation

SENTENCE$_1$: I like apples

SENTENCE$_2$: Apples are good for you

Vocabulary	Apples	Are	For	Good	I	Like	you
Sentence$_1$	1	0	0	0	1	1	0
Sentence$_2$	1	1	1	1	0	0	1

Similarity Computation: Cosine Measure

Cosine of angle between two vectors in n-dimensional space

$$sim(b_1, b_2) = \frac{\sum_t w_{y,b_1} w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_{t=1}^{n} w_{t,b_2}^2}}$$

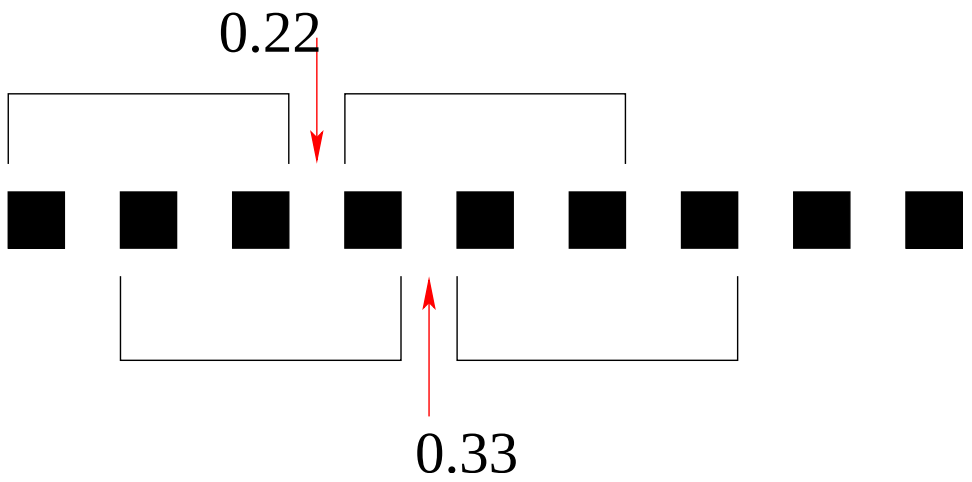
SENTENCE$_1$: 1 0 0 0 1 1 0

SENTENCE$_2$: 1 1 1 1 0 0 1

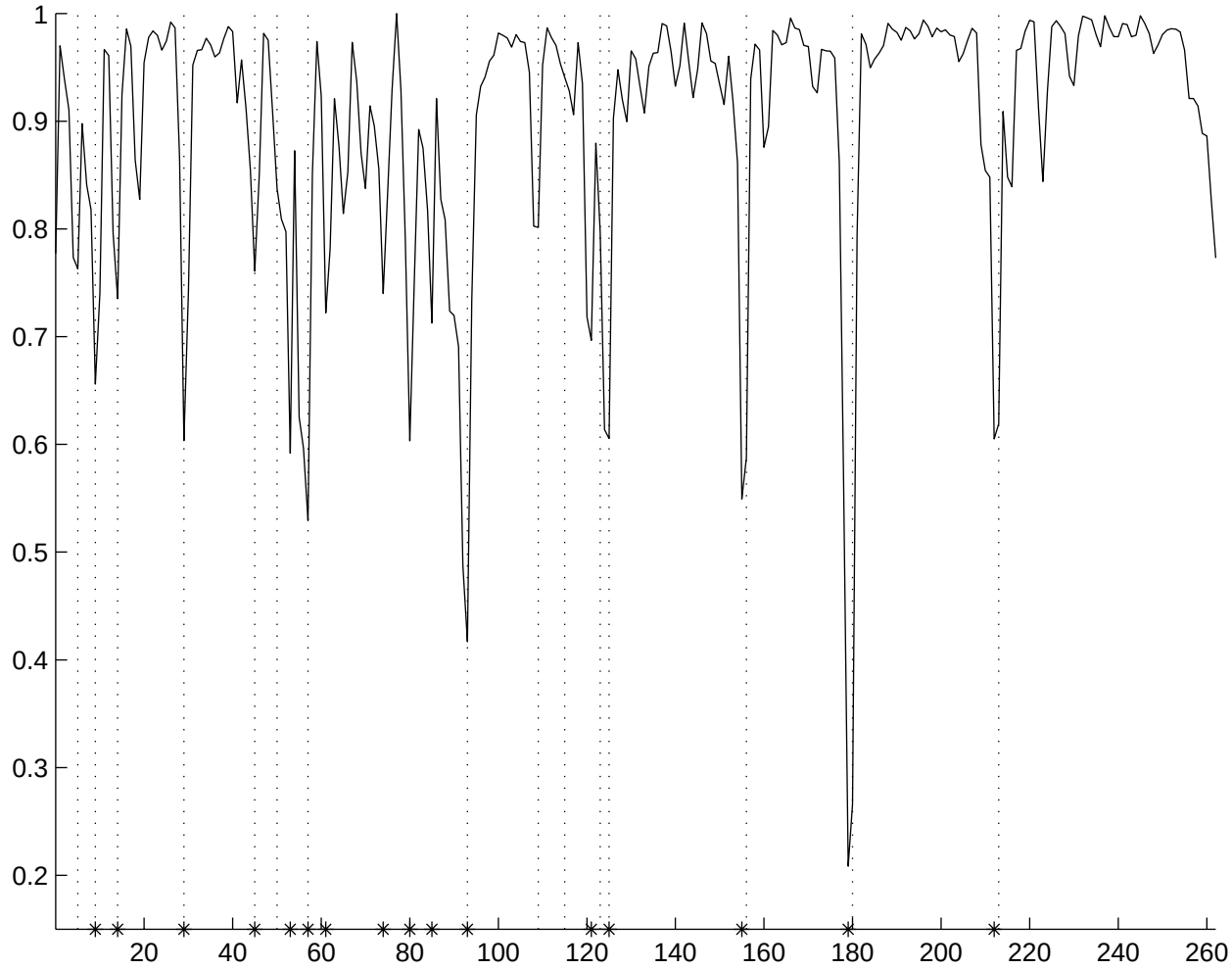
sim(S$_1$,S$_2$) =

$$\frac{1*0+0*1+0*1+0*1+1*0+1*0+0*1}{\sqrt{(1^2+0^2+0^2+0^2+1^2+1^2+0^2)*(1^2+1^2+1^2+1^2+0^2+0^2+1^2}} = 0.26$$

Similarity Computation: Output



Gap Plot


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
20
40
60
80
100
120
140
160
180
200
220
240
260

Boundary Detection

Based on changes in sequence of similarity scores:

Depth Scores: relative depth (in comparison to the closest maximum)

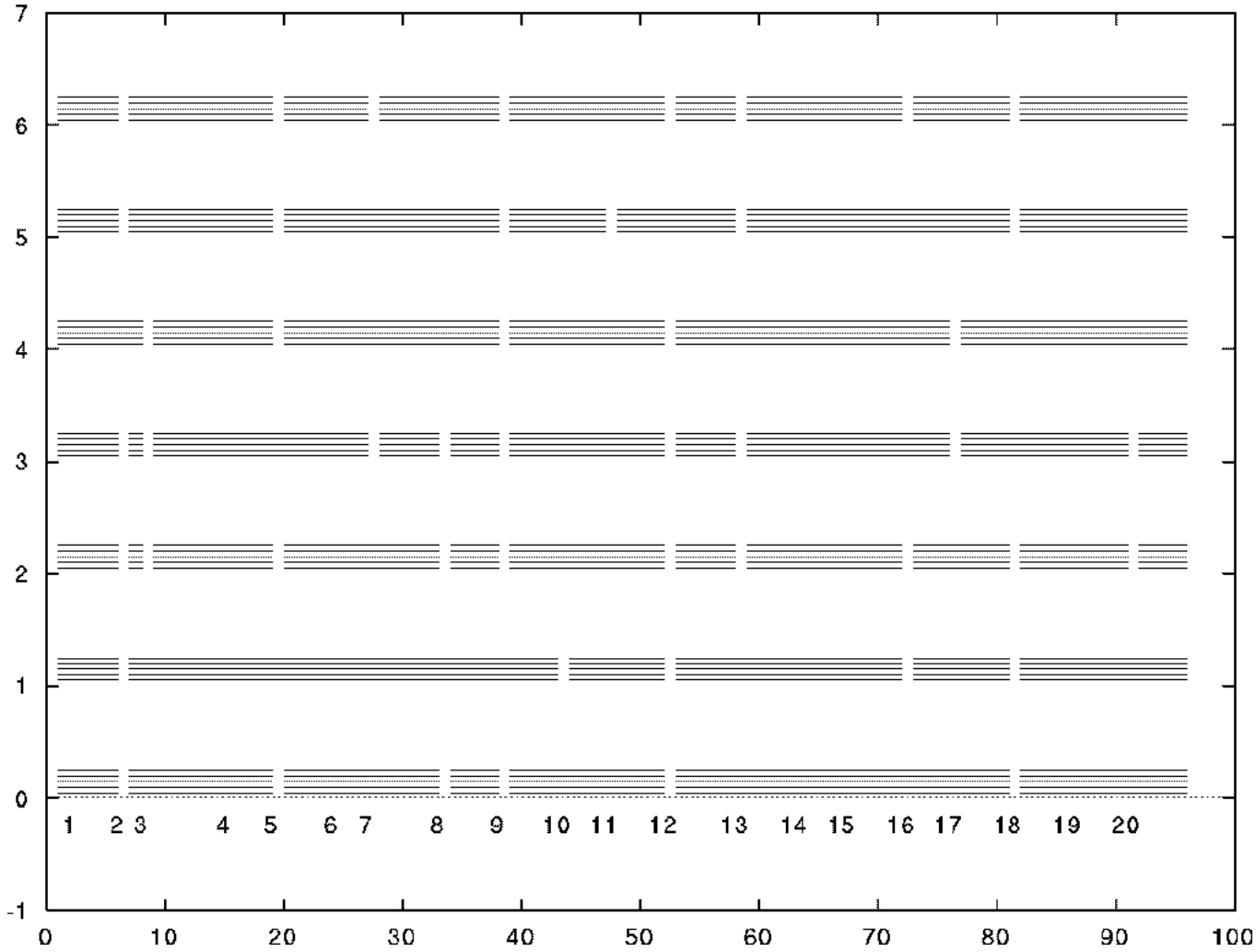
Number of segments: $s - \sigma/2$

Segmentation Evaluation

Comparison with human-annotated segments(Hearst'94):

- 13 articles (1800 and 2500 words)
- 7 judges
- boundary if three judges agree on the same segmentation point

Agreement on Segmentation



Evaluation Results

Methods	Precision	Recall
Baseline 33%	0.44	0.37
Baseline 41%	0.43	0.42
Chains	0.64	0.58
Blocks	0.66	0.61
Judges	0.81	0.71

More Results

- High sensitivity to change in parameter values
- Thesaural information does not help
- Most of the mistakes are “close misses”

Meeting Segmentation

- Motivation: Facilitate information Access
- Challenges:
  - High error rate in transcription
  - Multi-thread structure

Algorithm for Feature Segmentation

Supervised ML
(Galley&McKeown&Fosler-Lussier&Jing'03)

- Combines multiple knowledge source:
  - cue phrases
  - silences
  - overlaps
  - speaker change
  - lexical cohesion
- Uses probabilistic classifier (decision tree) to combine them

Cue Word Selection

Automatic computation of cue words:

- Compute word probability to appear in boundary position
- Select words with the highest probability
- Remove non-cues.

Selected Cue Words

OKAY	93.05
shall	0.44
anyway	0.43
alright	0.64
let's	0.66
good	0.81

Silences

- Pauses — speaker silence in the middle of her speech
- Gap — silences not attributable to any party

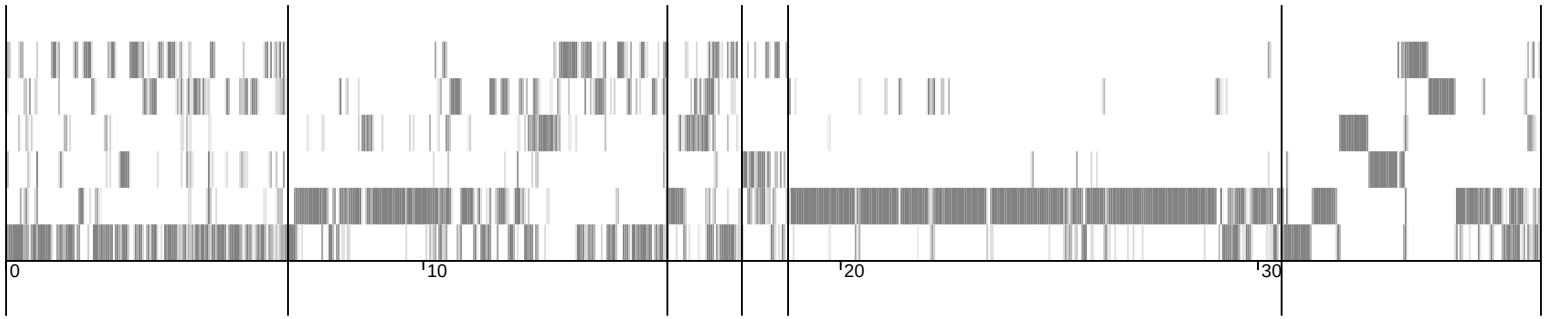
Topic boundaries are typically preceeded by gaps

Overlaps

- Average overlap rate within some window

Little overlap in the beginning of segments

Speaker Change



Determination of Window Size

Feature	Tag	Size(sec)	Side
Cue phrases	CUE	5	both
Silence (gaps)	SIL	30	left
Overlap	OVR	30	right
Speaker activity	ACT	5	both
Lexical cohesion	LC	30	both

Examples of Derived Rules

Condition	Decision	Conf.
LC$\leq$0.67, CUE$\geq$1, OVR$\leq$1.20, SIL$\leq$3.42	yes	94.1
LC$\leq$0.35, SIL$>$3.42, OVR$\leq$4.55	yes	92.2
CUE$\geq$1, ACT$>$0.1768, OVR$\leq$1.20, LC$\leq$0.67	yes	91.6
...		
default	no	

Results

Method	P_k	WD
Feature-based	23.00	25.47
Cohesion-based	31.91	35.88