Human vs. Machine

A	B
In December 1988 a Pan Am jet was blown up over Lockerbie, Scotland, killing 270. Since 1992 Libya has been under U.N. sanctions in effect until the suspects are turned over to United States or Britain. In August 1998 United States and Britain proposed a Netherlands trial. Libya asked for guarantees that the suspects would be incarcerated in Libya. Kofi Annan planned a December 1988 Libyan trip to move negotiations.	Secretary-General Kofi Annan said Wednesday that he may travel to Libya next week in hopes of closing a deal to try two Libyan suspects in the Pan Am Lockerbie bombing. The sanctions, were imposed to force Libyan leader Moammar Gadhafi to turn the men over. Louis Farrakhan, the leader of a U.S. Muslim group, congratulated on his recovery from a hip injury.

Automatic Cohesion Assessment

Regina Barzilay

April 28, 2004

Linguistic Foundations: Cohesion

Cohesion: language devices that connect individual sentences into a unified whole

Cohesion devices: repetition, coreference, ellipsis

1. There was once a little girl and a little boy and a dog
2. And the sailor was their daddy
3. And the little doggy was white
4. And they like the little doggy
6. And they fed it
7. And they ran away
8. And then daddy had to go on a ship
9. And the children misssed 'em

Readability Models

- Goal: induce a model that can predict the degree of text "well-formedness"
- Applications: summarization, question-answering, machine-translation
  - Evaluation tool
  - Scoring mechanism in probabilistic generation

Linguistic Foundations: Centering

Certain connectivity patterns among text entities are characteristic of all well-formed texts(Grosz, Joshi&Weinstein, 1995)

- Unit of analysis: centers
- "Affiliation" of a center: utterance (U) and discourse segment (DS)
- Function of a center: to link between a given utterance and other utterances in discourse

Center Typology

- Types:
  - Forward-looking Centers C_f (U, DS)
  - Backward-looking Centers C_b (U, DS)
- Connection: C_b (U_n) connects with one of C_f (U_{n-1})

Center Continuation

Continuation of the center from one utterance not only to the next, but also to subsequent utterances

- $C_b(U_{n+1})=C_b(U_n)$
- $C_b(U_{n+1})$ is the most highly ranked element of $C_f(U_{n+1})$ (thus, likely to be $C_b(U_{n+2})$

Center Shifting

Shifting the center, if it is neither retained no continued

- $C_b(U_{n+1}) <> C_b(U_n)$

Discussion on Centering

- Until now: always based on manual annotations
- Never used in applications

Does it really work?

Coherent Discourse

Coherence is established via center continuation

John went to his favorite music store to buy a piano.	John went to his favorite music store to buy a piano.
He had frequented the store for many years.	It was a store John had frequented for many years.
He was excited that he could finally buy a piano.	He was excited that he could finally buy a piano.
He arrived just as the store was closing for the day.	It was closing just as John arrived.

Entity matrix

1. [Former Chilean dictator Augusto Pinochet]$_S$, was arrested in [London]$_X$ on [14 October]$_X$ 1998.
2. [Pinochet]$_S$, 82, was recovering from [surgery]$_X$.
3. [The arrest]$_S$ was in [response]$_X$ to [an extradition warrant]$_X$ served by [a Spanish judge]$_O$.
4. [Pinochet]$_S$ was charged with murdering [thousands]$_O$, including many [Spaniards]$_O$.
5. [Pinochet]$_S$ is awaiting [a hearing]$_O$, [his fate]$_X$ in [the balance]$_X$.
6. [American scholars]$_S$ applauded the [arrest]$_O$.

	Dictator	Augusto	Pinochet	London	October	Surgery	Arrest	Extradition	Warrant	Judge	Thousands	Spaniards	Hearing	Fate	Balance	Scholars	
1	S	S	S	X	X	–	–	I	–	–	–	–	–	–	–	–	1
2	–	–	S	–	–	X	–	–	–	–	–	–	–	–	–	–	2
3	–	–	–	–	–	–	S	X	X	O	–	–	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	O	O	–	–	–	–	4
5	–	–	S	–	–	–	–	–	–	–	–	–	O	X	X	–	5
6	–	–	–	–	–	–	O	–	–	–	–	–	–	–	–	S	6

Our Approach

1. Construct an entity matrix that encodes distributional and syntactic information
2. Identify matrix patterns characteristic of well-formed texts

Computation of Entity matrix

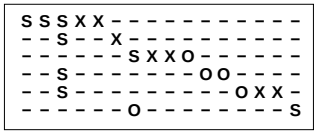
- Approximate discourse entities with nouns
- Mark all the members of noun compound with the same syntactic tag

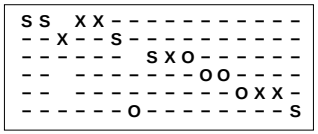
Matrix Properties

- Dense vs. sparse columns
- Distribution of syntactic tags

	HRS	MRS	LRS
s s	0.020	0.014	0.010
s o	0.012	0.005	0.004
- -	0.417	0.433	0.450

Matrix Comparison



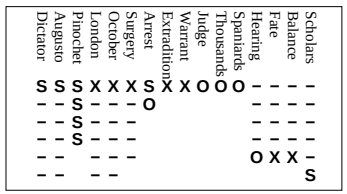


Transformations

Goal: reduce the variability in matrix representation

L	Original	Transformed
1.	- s - - - o - -	s - - - o - -
2.	- s - - - o - -	s - - - o
3.	- s - - - o - -	s o

Example of Transformation



Vector-Based Representation

s - - - - -	2
s s s s	1
x - - - - -	2
s o	1
x	2
o	3
- - - - o	1
- - - - x	2
- - - - - s	1

Discriminative Model

- Each document is described by fixed set of attributes (entity sequences) and their values (frequency)

Texts	x - - -	x s s	x - o	s - -	- o o
T_1	4	1	1	2	0
T_2	10	0	0	5	0
T_3	1	3	0	1	2

- Given a set of "good" and "bad" texts, we can learn attributes with high predictive power

Generative Model

- Probability of a column is defined using an n-gram model
- The probability of a matrix is computed by multiplying together the probability for each column and normalizing the product

Experiments: Data

Humans	HRS	MRS	LRS
5.13	4.42	4.32	3.60

Results of Anova Analysis:

- Human summaries are more cohesive than machine generated ones
- HRS is not significantly different from MRS
- Both HRS and MRS are significantly more cohesive than LRS

Experiments: Data

- Data: Outputs of three multi-document summarization systems that participated in DUC'2003 and corresponding human summaries
  - High grammaticality scores
  - Variability in readability scores: High (HRS), Medium (MRS) and Low (LRS)
  - Overall 64 summaries
- Procedure: the judge assigns readability score on a seven point scale
  - 183 summaries (23 people per summary)

Model Comparison: Baselines

- Readability Measures — a function of the average sentence length and the average number of syllables (Flesh, 1951)
- Word-based Models — the average word overlap of adjacent sentences (Foltz&Kintsch&Landauer, 1998)
- Vector-based Models — the average distances between adjacent sentences based on word distributional properties (Foltz&Kintsch&Landauer, 1998)
- Taxonomy-based Models — the average distances between adjacent sentences based on WordNet (Lin, Resnik)

Agreement

Why not to use kappa?

- Function: Upper-bound on human performance
- Procedure: Leave-one-out resampling (Weiss&Kulikowski)
- Result: Agreement = .612 (Min = .107, Max = .975, SD = .230)

Generative Model: Implementation

- Applied to 6-letter alphabet at various level of compression
- Trained on DUC human summaries
- Tested on machine summaries

Results: Generative Model

Correlation between human rating and the models

Model	Correlation
Flesh Readability Index	.010
Word-based Model	.113
Latent Semantic Analysis	.184
Taxonomy-based (Lin)	$-.125$
Taxonomy-based (Resnik)	$-.176$
Entity Matrix	$.314^{**}$
$^{*}p < .05$ (2-tailed)	
$^{**}p < .01$ (2-tailed)	

Discussion: Generative Model

- No correlation for traditional cohesion model due to redundancy
- High negative correlation for Wordnet-based models!
- Best results on the tranformation 3

Results: Discriminative Model

Trans	Vec. Size	2-way	3-way
Base	-	69%	37.5%
0	354	73%	53.3%
1	101	97%	59.4%
2	88	97%	59.4%
3	77	97%	64.1%
4	73	78%	64.1%
N	56	73%	62.5%

Future directions

- Dependence on genre
- Contribution of different linguistic features
  - Preliminary results: anaphora doesn't help
- More sophisticated model (unsupervised grammar induction, gap modeling)

Discussion: Discriminative Model

- Most predictive patterns: `[s x]`, `[x o]`, `[s - s]` and `[s s s]`
- Baselines: binary 67%, trinary 37,5%
- Transformation 3 is optimal in all the cases