

Text Quality Assessment

Sources of text

Alignment

- User generated
 - Articles
 - Blogs
 - News
 - Literary texts
 - Scientific Articles
 - Reviews
- Machine Generated
 - Language models
 - Machine Translators
 - Summarizers

What is Quality

- “the standard of something as measured against other things of a similar kind; the degree of excellence of something.”
- “a distinctive attribute or characteristic possessed by someone or something.”

How Quality of a Text Measured

- Information Extraction
- Product Review Analysis
- Summarization
- Language Understanding
- Machine Translation
- Answer Evaluation
- ...



- Relevance
- Completeness
- Informativeness
- Readability
- Well-formedness
- Cohesion
- Lexical diversity
- Grammar

How Quality of a Text Measured

- Information Extraction
- Product Review Analysis
- Summarization
- Language Understanding
- Machine Translation
- Answer Evaluation
- ...



- Relevance
- Completeness
- Informativeness
- Readability
- Well-formedness
- Cohesion
- Lexical diversity
- Grammar

Text Readability

Readability is “... *the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success of a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting*” (Edgar Dale and Jean Chall, 1949)

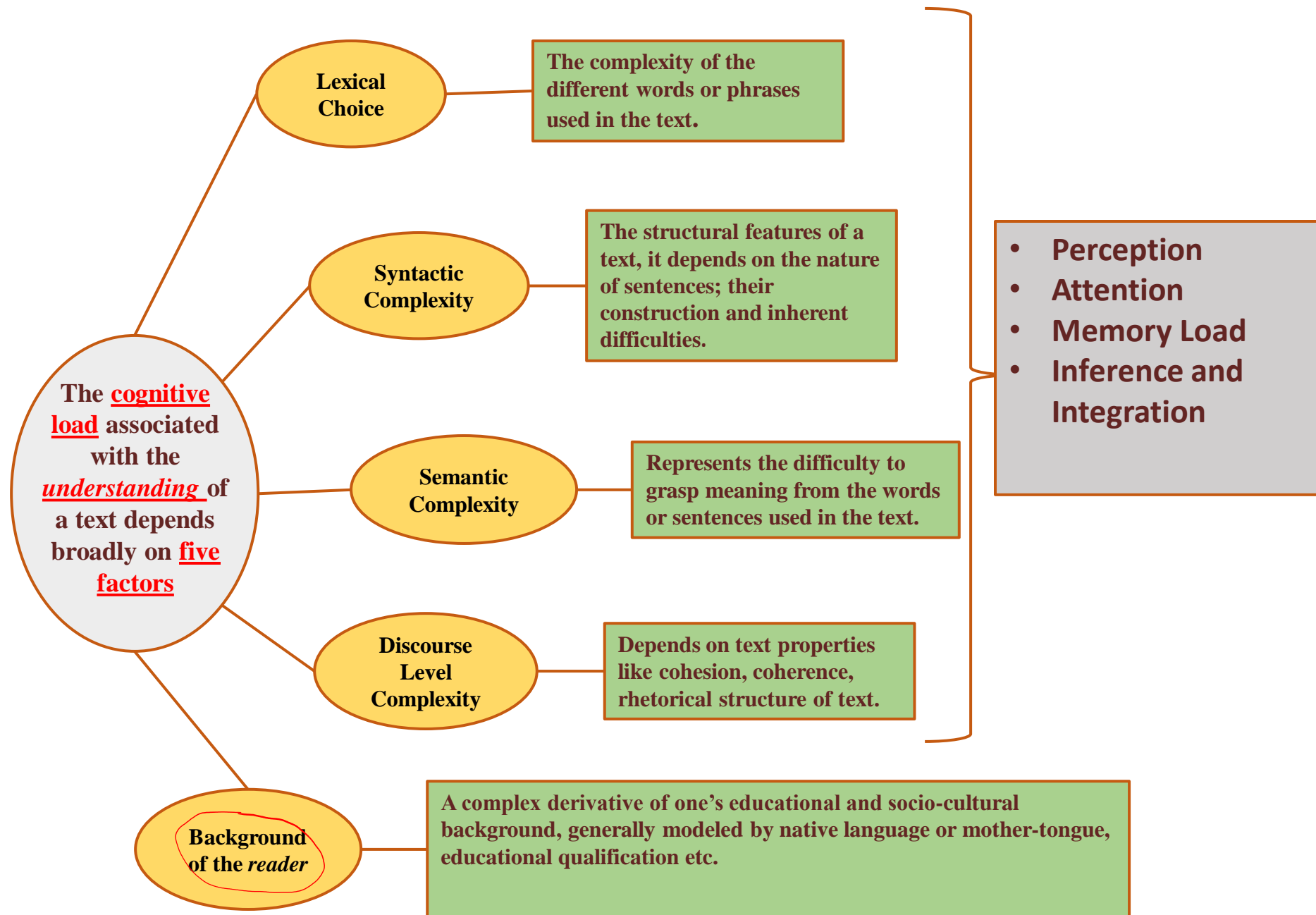
or to put it simply: The factors that contributes to the **cognitive load** (of the target reader) associated with comprehending a text.

Excerpts on Light from Wikipedia:

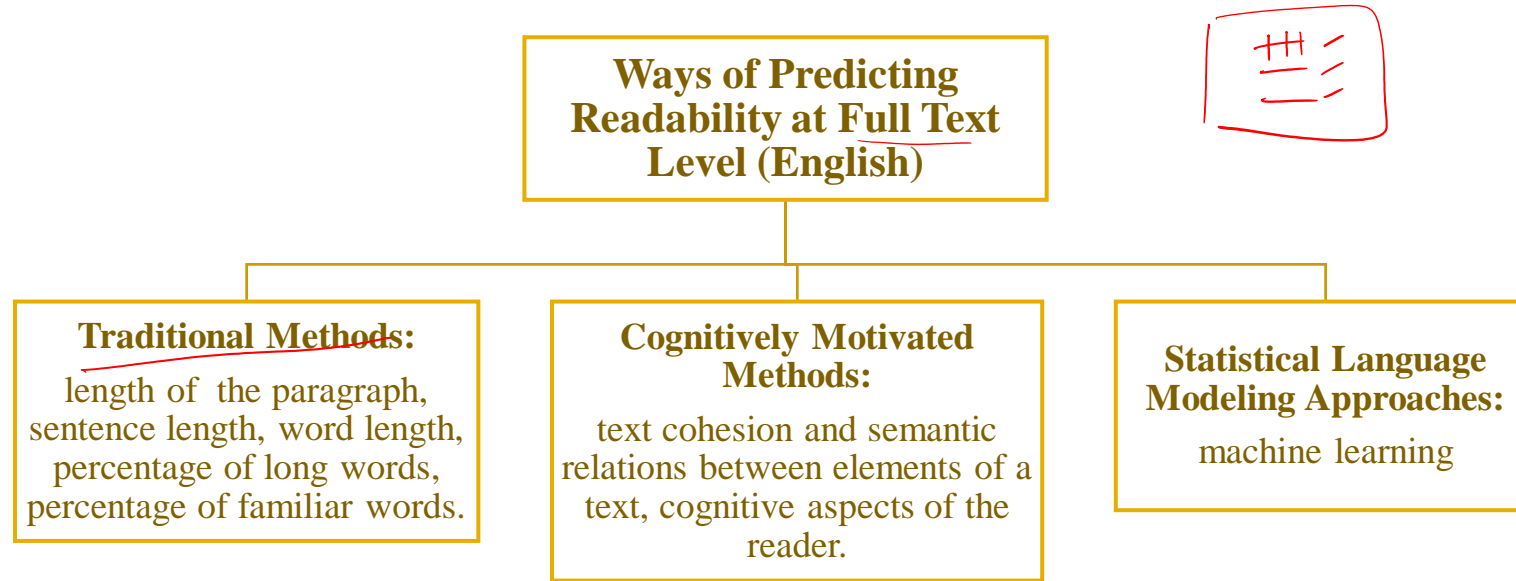
“Light usually refers to visible light, which is electromagnetic radiation that is visible to the human eye and is responsible for the sense of sight. Visible light is usually defined as having a wavelength in the range of 400 nanometres (nm), or 400×10^{-9} m, to 700 nanometres – between the infrared, with longer wavelengths and the ultraviolet, with shorter wavelengths....”

Excerpts on Light from Wikipedia Kids:

“Light is a type of energy. It is a form of electromagnetic radiation of a wavelength which can be detected by the human eye. It is a small part of the electromagnetic spectrum and radiation given off by stars like the sun. Animals can also see light. Light exists in tiny packets called photons. It shows properties of both waves and particles. The study of light, known as optics, is an important research area in modern physics.....”



The language context: Readability formulae in different languages



- Similar techniques used for languages other than English
 - French, Spanish, Swedish, Italian, Thai, Arabic, Portuguese, Danish
- Formulae tuned according to language specific constraints

Readability Formulas: Flesch-Kincaid Grade Level

- Designed to indicate how difficult a passage in English is to understand.
- There are two tests: the Flesch Reading-Ease, and the Flesch–Kincaid Grade Level.
- Although they use the same core measures (word length and sentence length), they have different weighting factors.

$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

Readability Formulas: Flesch-Kincaid Grade Level

| Score | School level (US) | Notes |
|--------------|-------------------------------------|---|
| 100.00–90.00 | 5th grade | Very easy to read. Easily understood by an average 11-year-old student. |
| 90.0–80.0 | 6th grade | Easy to read. Conversational English for consumers. |
| 80.0–70.0 | 7th grade | Fairly easy to read. |
| 70.0–60.0 | 8th & 9th grade | Plain English. Easily understood by 13- to 15-year-old students. |
| 60.0–50.0 | 10th to 12th grade | Fairly difficult to read. |
| 50.0–30.0 | College | Difficult to read. |
| 30.0–10.0 | College graduate | Very difficult to read. Best understood |
| 10.0–0.0 | Professional | Extremely difficult to read. Best unders |

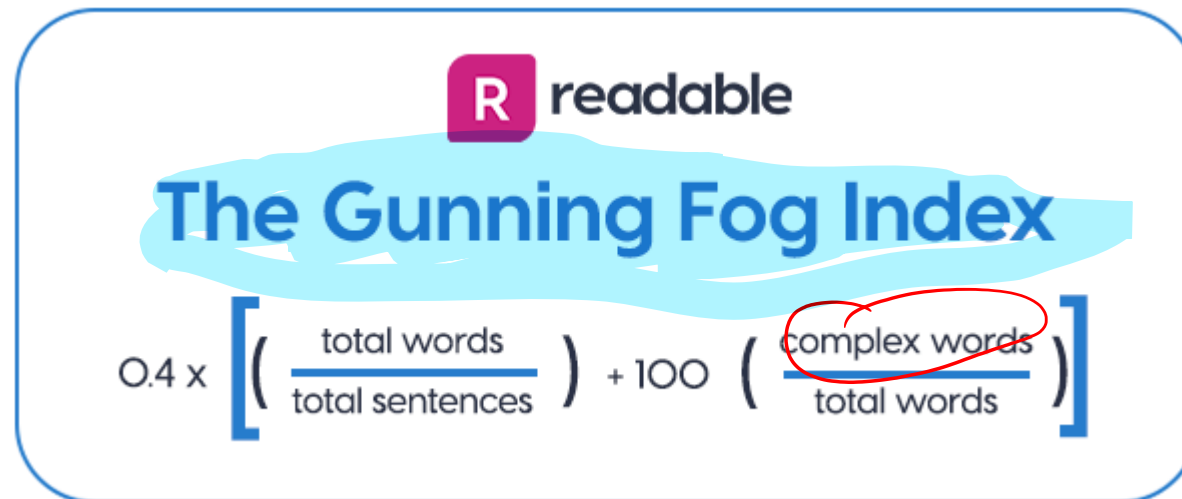
[Reader's Digest](#) magazine has a readability index of about 65, [Time](#) magazine scores about 52, an average grade six student's written assignment (age of 12) has a readability index of 60–70

The average Flesch score is 49.13, the highest score (81.32) is for [Stone](#) and the lowest is for [Order of the Phoenix](#)



Readability Formulas: Gunning-Fog Score

- Select a passage (such as one or more full paragraphs) of around 100 words. Do not omit any sentences;
- Determine the average sentence length. (Divide the number of words by the number of sentences.);
- Count the "complex" words consisting of three or more syllables. Do not include proper nouns, familiar jargon, or compound words. Do not include common suffixes (such as -es, -ed, or -ing) as a syllable;
- Add the average sentence length and the percentage of complex words; and
- Multiply the result by 0.4.

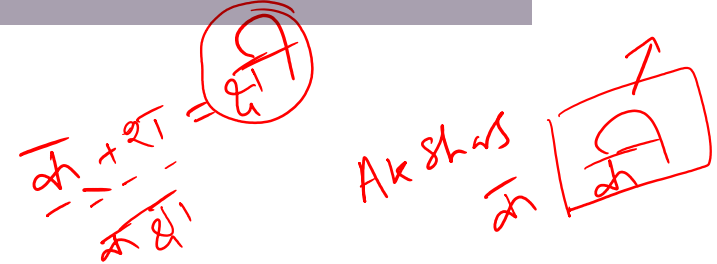


The Gunning Fog Index

$$0.4 \times \left[\left(\frac{\text{total words}}{\text{total sentences}} \right) + 100 \left(\frac{\text{complex words}}{\text{total words}} \right) \right]$$

Readability Formulas: Coleman-Liau Index

$$CLI = 0.0588L - 0.296S - 15.8$$



- L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

Existing computer programs that measure readability are based largely upon subroutines which estimate number of syllables, usually by counting vowels. The shortcoming in estimating syllables is that it necessitates keypunching the prose into the computer. There is no need to estimate syllables since word length in letters is a better predictor of readability than word length in syllables. Therefore, a new readability formula was computed that has for its predictors letters per 100 words and sentences per 100 words. Both predictors can be counted by an optical scanning device, and thus the formula makes it economically feasible for an organization such as the U.S. Office of Education to calibrate the readability of all textbooks for the public school system.

$$L = \text{Letters} \div \text{Words} \times 100 = 639 \div 119 \times 100 \approx 537$$

$$S = \text{Sentences} \div \text{Words} \times 100 = 5 \div 119 \times 100 \approx 4.20$$

$$CLI = 0.0588 \times 537 - 0.296 \times 4.20 - 15.8 = 14.5$$

Therefore, the abstract is at a grade level of 14.5, or roughly appropriate for a second-year undergraduate.

Readability Formulas: SMOG Index

1. Count a number of sentences (at least 30)
2. In those ~~sentences~~, count the polysyllables (words of 3 or more syllables).
'sad' *(elephant)*
3. Calculate using

$$\text{grade} = 1.0430 \sqrt{\text{number of } \underline{\text{polysyllables}} \times \frac{30}{\text{number of sentences}}} + 3.1291$$

This version (sometimes called the SMOG Index) is more easily used for mental math:

1. Count the number of polysyllabic words in three samples of ten sentences each.
2. Take the square root of the nearest perfect square
3. Add 3

3 with fish
Beauty + fish
↓

Limitations

- They neglect between-reader differences and effects of content, layout and retrieval aids
- Not all complex words are difficult. For example, "interesting" is not generally thought to be a difficult word, although it has three syllables.
- tables for texts of fewer than 30 sentences are statistically invalid

Reading difficulty of full Text

Submitted
Text Answer
Assessment

Excerpts on Light from Wikipedia:

"Light usually refers to visible light, which is electromagnetic radiation that is visible to the human eye and is responsible for the sense of sight. Visible light is usually defined as having a wavelength in the range of 400 nanometres (nm), or 400×10^{-9} m, to 700 nanometres – between the infrared, with longer wavelengths and the ultraviolet, with shorter wavelengths...."

Readability Statistics:

- Flesch-Kincaid Grade Level: 10.8
- Gunning-Fog Score: 14.4
- Coleman-Liau Index: 13.6
- SMOG Index: 10.5

Text Statistics

- Characters per Word: 5.0
- Syllables per Word: 1.7
- Words per Sentence: 16.5

Excerpts on Light from Wikipedia Kids:

"Light is a type of energy. It is a form of electromagnetic radiation of a wavelength which can be detected by the human eye. It is a small part of the electromagnetic spectrum and radiation given off by stars like the sun. Animals can also see light. Light exists in tiny packets called photons. It shows properties of both waves and particles. The study of light, known as optics, is an important research area in modern physics...."

Readability Statistics:

- Flesch-Kincaid Grade Level: 8.2
- Gunning-Fog Score: 11.2
- Coleman-Liau Index: 11.2
- SMOG Index: 8.1

Text Statistics

- Characters per Word: 4.6
- Syllables per Word: 1.6
- Words per Sentence: 12.1

Modeling word level difficulty

- ❑ Visual word recognition (in *isolation* has only been considered here) has been a field of constant investigation.
- ❑ Complexity of a word can be attributed to **4 factors**:
 - **Orthography** : how the word looks?
 - **Phonology** : how it sounds? (Example: **Psychology**, **Pneumonia**)
 - **Morphology** : **Beauty** -> **Beautification**, **Help** -> **Helplessness**
 - **Semantics** : which concept (s) it represents? (Example (# no. of different meanings: **Run (179)**, **Turn (122)**, **Play (95)**)
 - **Compound words**: (Example: **Birthday**, **Pickpocket**, **Brunch**)
 - **Multiword expressions**: (Example: **Swimming-pool**, **Cloud-nine**)

Category of features

Features considered:

❑ 6 syntactic features:

- average sentence length
- Average word length in terms of visual units
- Average word length in syllable
- No. of polysyllabic words
- No. Of polysyllabic words in 30 sentences
- No. of consonant conjuncts/jukta-akshars in 50 sentences

❑ 10 POS features

❑ Sentence level

- \$(noun phrase)
- \$(verb phrase)
- \$(adjective)
- \$(postposition)
- \$(entity)
- \$(unique entity)
- \$(clauses)

❑ Discourse level

- \$(noun phrase)
- \$(verb phrase)
- \$(adjective)
- \$(postposition)
- \$(entity)
- \$(unique entity)

❑ 2 discourse cohesion features

- number of lexical chain
- average lexical chain length

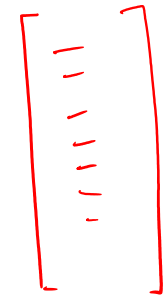
Analysis procedures: Three different combination of features were tested for effectiveness

1. Only syntactic
2. Only POS and discourse
3. Both

What is Syntax

- **Structure of language**
 - How words are arranged together and related to one another
 - Ordering words in sequences to express meanings for which no separate word exists.
 - **Goal of syntactic analysis**
 - relate surface form to underlying structure, to support semantic analysis
 - **Two views of linguistic structure:**
 1. **Constituency** = phrase structure grammar = context-free grammars (CFGs)
 2. **Dependency Structure**
 - Dependency structure shows which words depend on (modify or are arguments of) which other words.
- **Starting unit: words** are given a category (part of speech = pos)
the, cat, cuddly, by, door
Det N Adj P N
 - **Words combine into phrases** with categories
the cuddly cat, by the door
NP → Det Adj N PP → P NP
 - **Phrases can combine into bigger phrases** recursively
the cuddly cat by the door
NP → NP PP
- Many aspects of meaning can be learnt using the syntactic structure.
 - The NP preceding VP is likely the subject of the action.
 - The NP following the VP is likely the object of the action.
 - Knowing basic units is helpful in modeling language.
 - You can use this to predict or complete the sentence.
 - Re-organize sentences or simplify them.

What is Dependency Parsing?

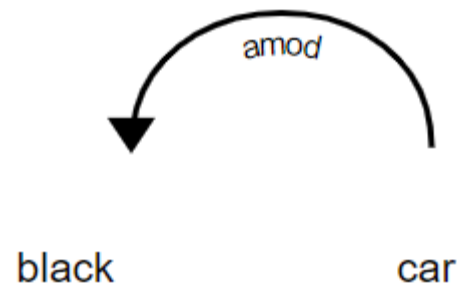


- Dependency Parsing is the process to analyze the grammatical structure in a sentence and find out related words as well as the type of the relationship between them.

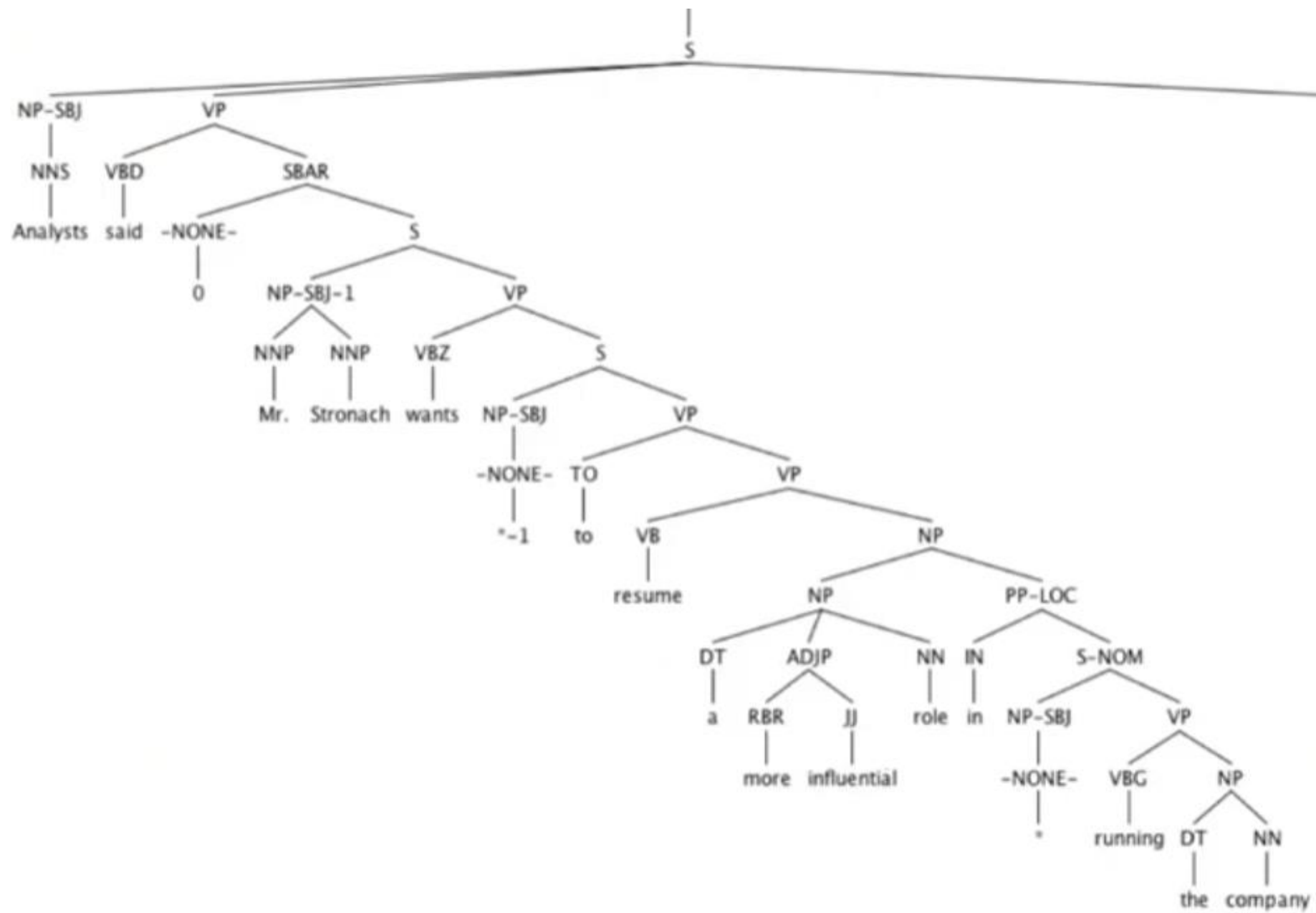
Each relationship:

1. Has one **head** and a **dependent** that modifies the **head**.
2. Is labeled according to the nature of the dependency between the **head** and the **dependent**. These labels can be found at [Universal Dependency Relations](#).

-



Simple dependency relation between two words



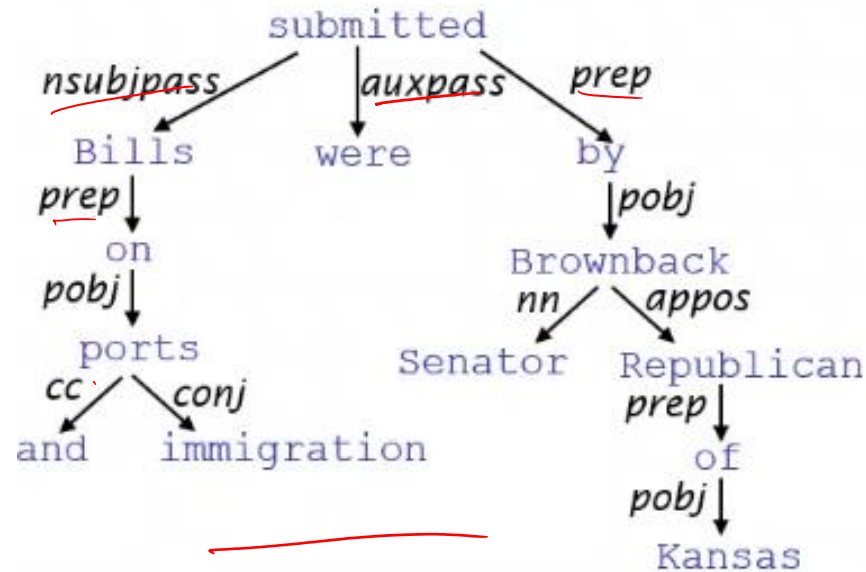
Phase structure grammar and Dependency grammar

Dependency relations: The dependency relations among words in sentences can be represented as a directed acyclic graph (DAG).

Simplified dependency parser

Bills on ports and immigration were submitted by Senator Brownback Republican of Kansas

The arrows are commonly **typed** with the name of grammatical relations



- nsubj:pass(submitted-7, Bills-1)
- case(ports-3, on-2)
- nmod(Bills-1, ports-3)
- cc(immigration-5, and-4)
- conj(ports-3, immigration-5)
- aux:pass(submitted-7, were-6)
- root(ROOT-0, submitted-7)
- case(Republican-11, by-8)
- compound(Republican-11, Senator-9)
- compound(Republican-11, Brownback-10)
- obl(submitted-7, Republican-11)
- case(Kansas-13, of-12)
- nmod(Republican-11, Kansas-13)

Dependency relations: Properties

- Using this graph some of the structural properties like, flatness, embeddedness, width of dependency, depth of dependency, average dependency distance and longest dependency distance are computed.
 - The **flatness** of a sentence is obtained by computing the average degree of each node (that represents a word) of a given DAG. The degree of a node/word represents how many words are dependent on that word.
 - **Embeddedness** is computed by closeness centrality or the count of number of nodes/words in between the root and a given node.
 - **Width of dependency**: Number of words that are dependent on one word
 - **Depth of dependency**: Number of words between the main verb and a given word.

Dependency distance

- The dependency distance between a pair of related words is computed by counting the number of words that occur between the two words.
- For example, in the sentence “I went to the market to buy a book” the dependency relations returned by the Stanford parser are

- nsubj(went-2, I-1)
- root(ROOT-0, went-2)
- case(market-5, to-3)
- det(market-5, the-4)
- obl(went-2, market-5)
- mark(buy-7, to-6)
- xcomp(went-2, buy-7)
- det(book-9, a-8)
- obj(buy-7, book-9)

The cat sat on the mat sitting on the mat was playing with the doll and the ball.

- The 1st line of the above representation indicates the second word “went” is dependent on the 1st word “I” and the type of relation is nSubj.
- From the above relations we compute the dependency distance between the word pairs “I” and “went” is 2-1=1. Therefore,
- The longest dependency distance in this sentence will be= $\forall_{ij} \max(\text{distance}(w_i, w_j))$
- Average dependency distance (ADD) of a sentence

$$= \frac{\text{Sum of distance of all the dependencies in the sentence}}{\text{\# of dependencies of the sentence}}$$

- Average dependency distance of a document

$$= \frac{\text{Sum of the ADD of the sentences}}{\text{\# of sentence of the document}}$$

Example: *Sarah read the book quickly and understood it correctly.*

(dependency distances calculated by Stanford Parser)

nsubj (read-2, Sarah-1)
det (book-4, the-1)
dobj (read-2, book-4)
advmod (read-2, quickly-5)
cc (read-2, and-6)
conj (read-2, understood-7)
dobj (understood-7, it-8)
advmod (understood-7, correctly-9)

Total dependency distance:

Number of dependency :

ADD =

Example: *Sarah read the book quickly and understood it correctly.*

(dependency distances calculated by Stanford Parser)

nsubj (read-2, Sarah-1) => 2-1 => 1

det (book-4, the-1) => 4-1 => 3

dobj (read-2, book-4) => 4-2 => 2

advmod (read-2, quickly-5) => 5-2 => 3

cc (read-2, and-6) => 6-2 => 4

conj (read-2, understood-7) => 7-2 => 5

dobj (understood-7, it-8) => 8-7 => 1

advmod (understood-7, correctly-9) => 9-7 => 2

Total dependency distance: (1+3+2+3+4+5+1+21)= 21

Number of dependency : 8

ADD = 21/8 = 2.62

Entropy based method

- Sentence processing occurs eagerly and incrementally.
- Producer and comprehender share the same grammar.
- “How much information does a comprehender get from a word?”= “how much uncertainty about the derivation reduced?”
- **Change in Entropy** is positively related to human sentence comprehension.

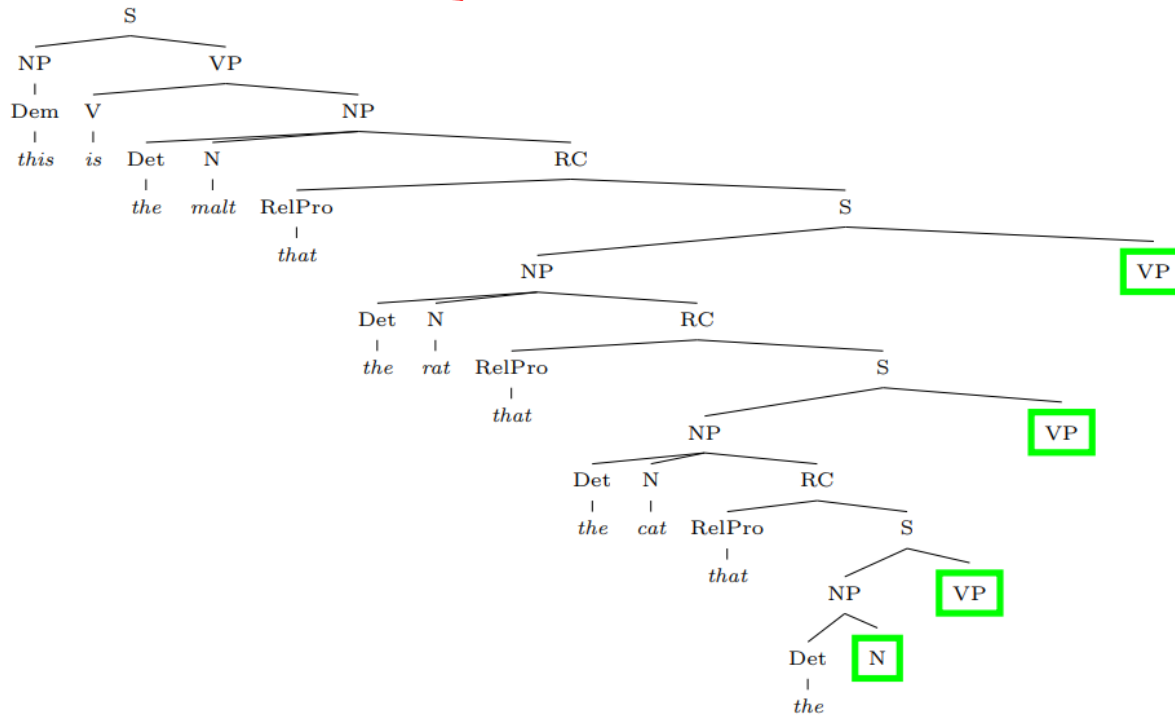
$$H(X) = -\sum_{x \in X} p_x \log p_x$$

(The old man the boat).
The old man was

w_1, w_2, w_3, \dots, p

Clausal Embeddings

- This is the malt that was eaten by the rat that was killed by the cat that was worried by the dog.

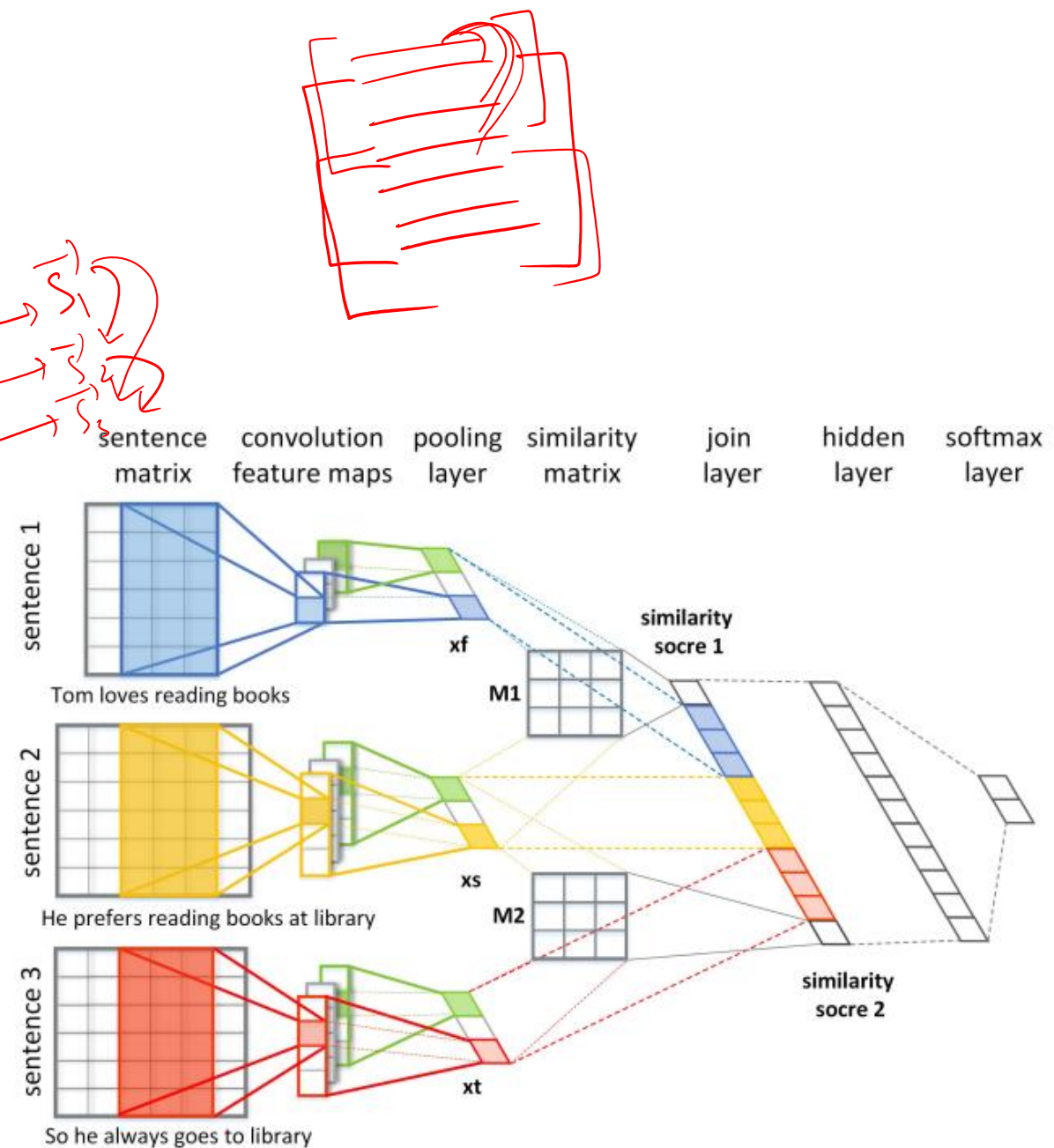
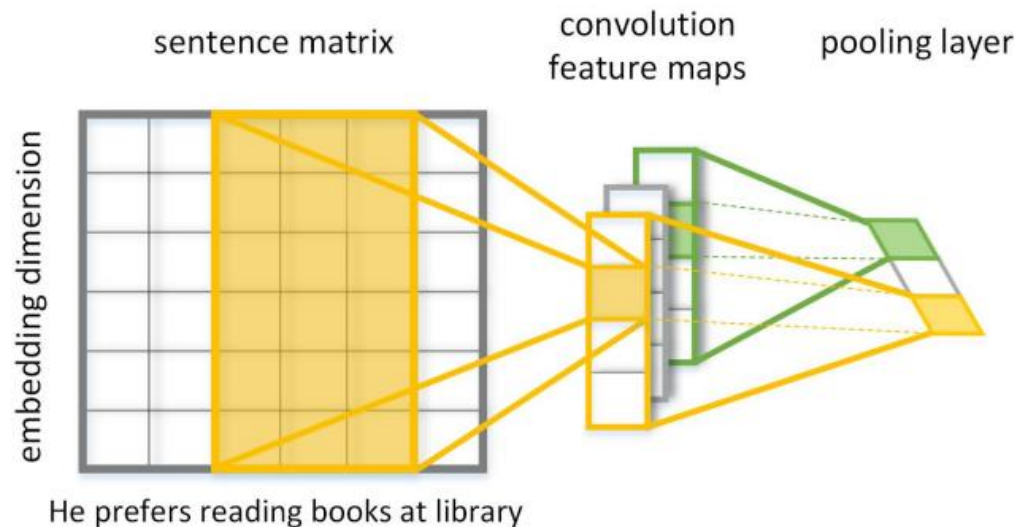


The motorcyclist who (the fact that the bike messenger blocked traffic) angered just missed the car.

Coherence

Refers to the semantic link between pair of text segments.

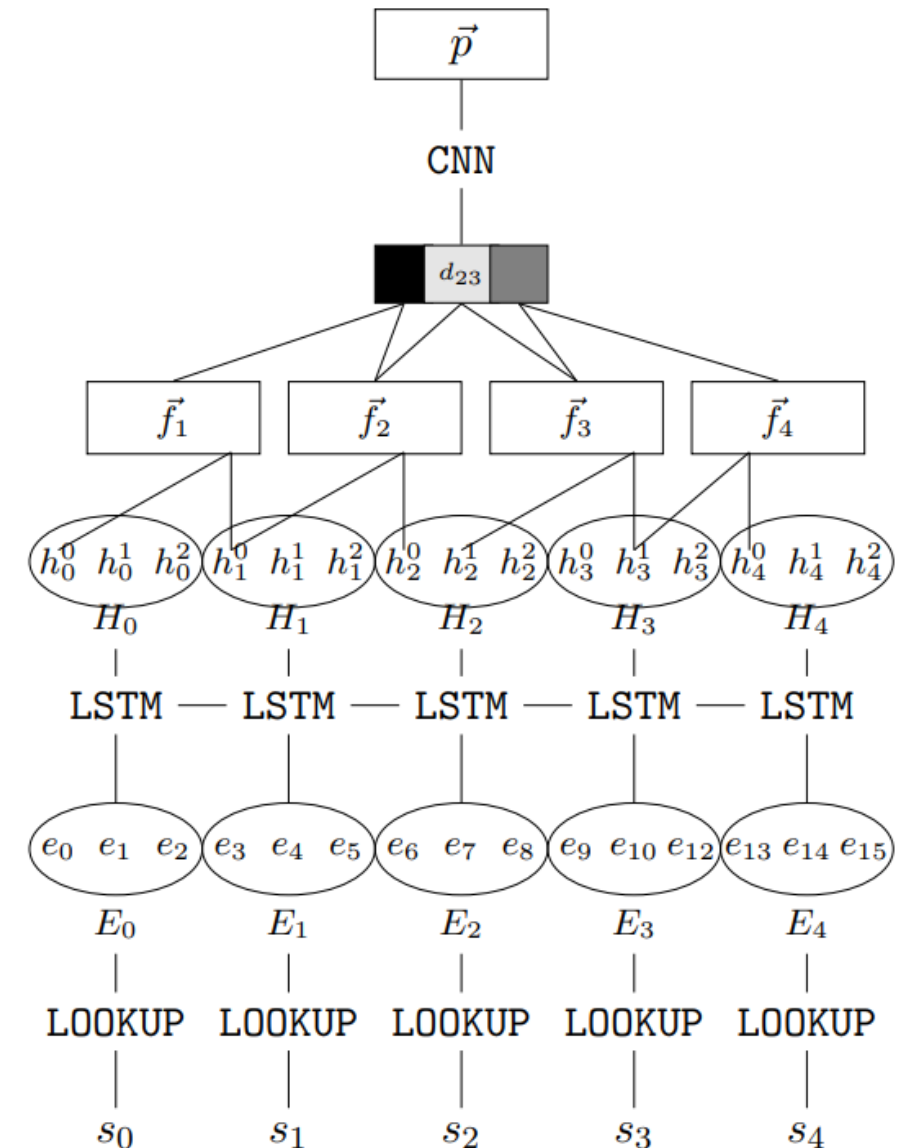
| Text 1 | Text 2 |
|---|---|
| Tom loves reading books. He prefers reading books at library. So he always goes to library. | Tom loves reading books. He missed his lunch today. So he always goes to library. |
| label=1 (coherent) | label=0 (incoherent) |



The architecture of the deep coherence model (DCM) for text coherence analysis with two matrices to encode similarity between adjacent sentences.

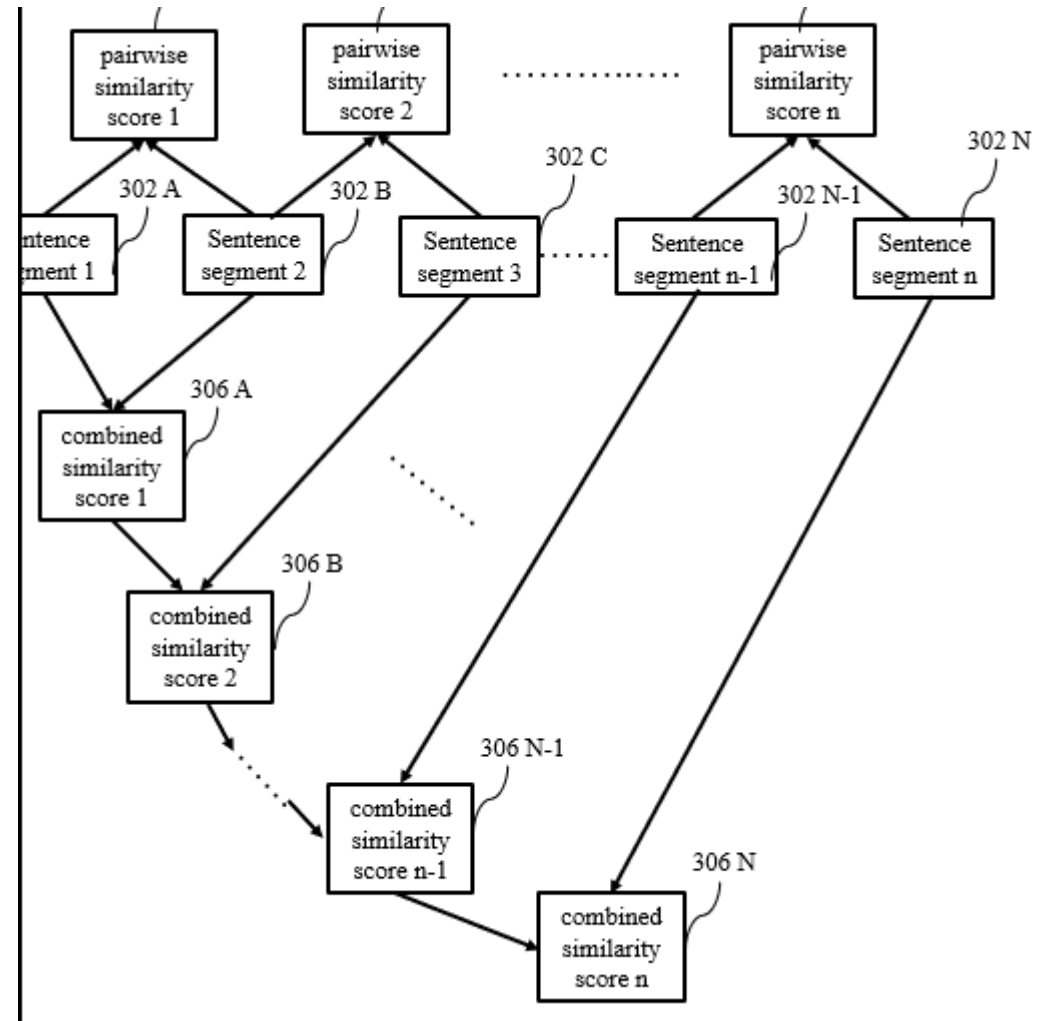
Coherence

- e_k is word embeddings associated with the k^{th} word in an input text.
- h_{ji} depicts the j th hidden state in LSTM states of sentence s_i .
- Two states in LSTM states of sentence s_i and sentence s_{i-1} that have the highest similarity are selected to connect sentences.
- Vector f_i captures information about the salient topic that relates sentence s_i to sentence s_{i-1} .
- d_{23} represents the similarity between f_2 and f_3 or the degree of continuity of the topic over adjacent sentences.
- Different shades of gray show different degrees of similarity.
- The CNN encodes patterns of changes as coherence vector p .



Coherence: Sliding Window Method

- The algorithm measures coherence by calculating semantic similarity for successive nonoverlapping segments of a text document.
- Each segment contains a predefined window size of x and y sentences.
- We represent the two segments as $S_{1,x}^{<i>}$ and $S_{2,y}^{<i+1>}$ with two different window size $x=1$ and $y=2$ respectively and i is the index of the segment at a given time step.
- For each time-step the algorithm selects segments $S_{1,x}^{<i>}$ and $S_{2,y}^{<i+1>}$ and compute the semantic similarity of the segments.
- For example, we begin with computing similarity between segment 1 and 2.
- Then, similarity is estimated for segment 2 and 3, then 3 and 4, and so on for the entire document.
- The final score is the average of the estimated similarities.
- The semantic similarity between two pair of segments is computed using a standard Bidirectional Encoding Representations for Transformers (BERT) based neural network architecture (Devlin et al., 2018).



Lexical Diversity

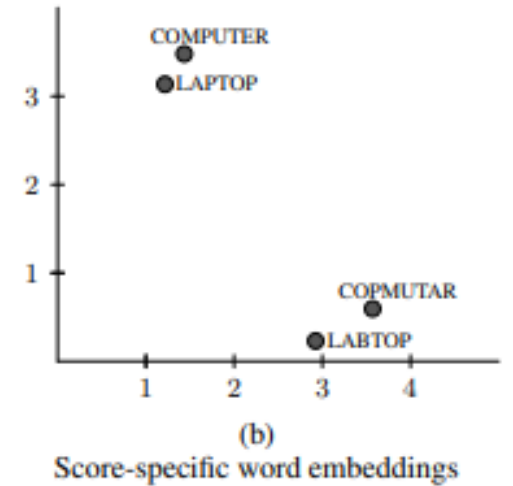
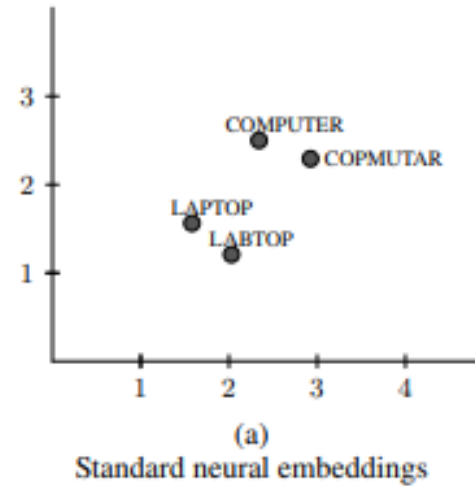
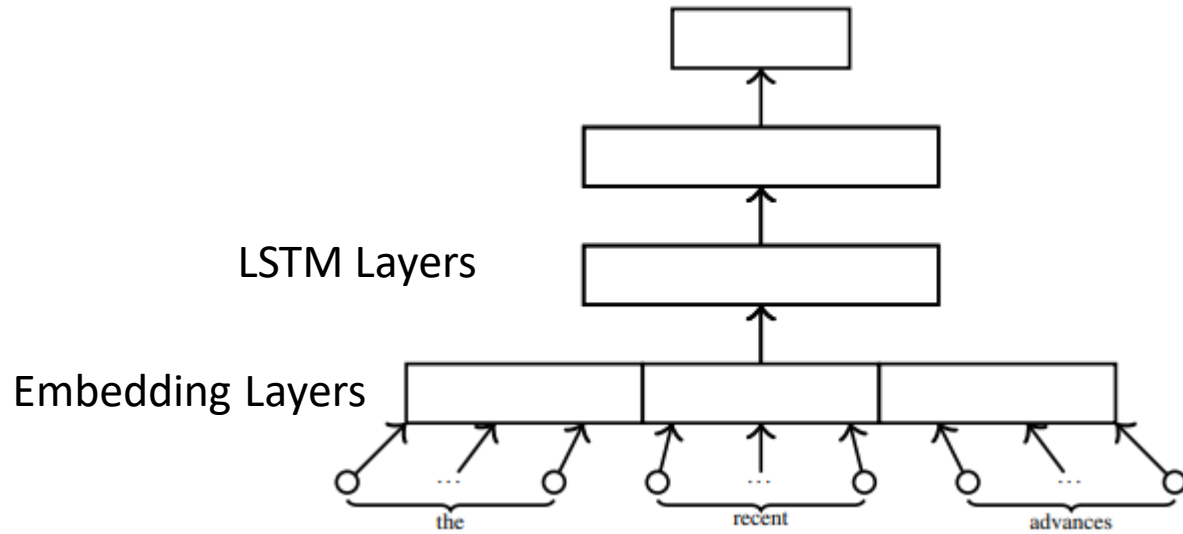
- Lexical diversity is a measure of how many different words appear in a text and it can be calculated in several different ways.
- A lexically diverse text is likely to describe the way the person cries using a variety of synonyms such as 'weep', 'shed tears', 'sob', 'wail', 'wimper', 'bawl', and so on. These words are more specific than if the person was to just use the vocabulary word 'cry' as they can express a range of emotions.
- Most of these techniques typically compute the ratio of unique word (types) to the total number of words (tokens) in a document, also known as a type-token ratio (TTR).
- However, all the existing techniques assume a lexical entity to be a single word and thus overlook the phrasal expressions like, multi-word expressions, complex predicates and compound verbs.
- We propose a novel technique that combines the standard TTR with phrasal density estimation.
- The phrasal density of a document is estimated as:

$$PD = \frac{\text{\# key phrases in a document}}{\text{Total No.of phrases}}$$

- Finally, we compute the overall lexical density as:

$$LD = TTR + PD$$

Application: Automatic Assessment of Text



Application: Automatic Assessment of Text

Linguistically Informed Networks

