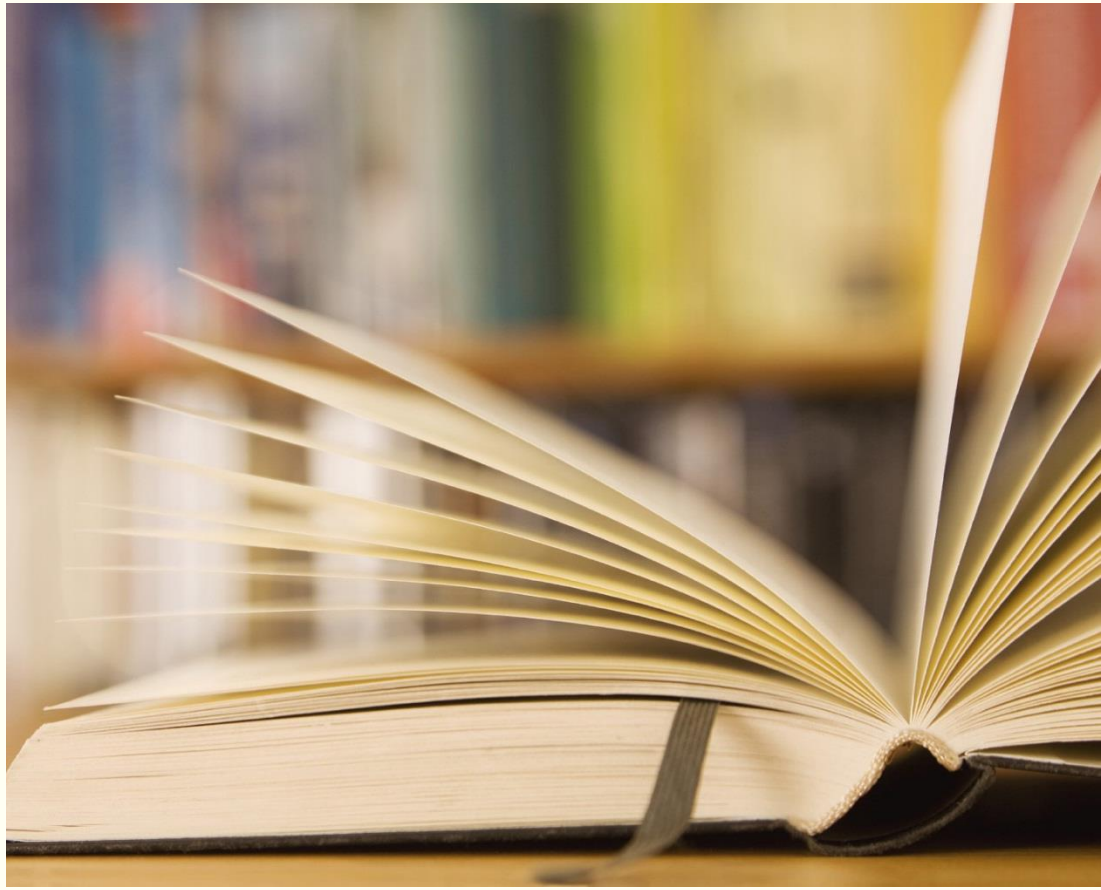
NATURAL LANGUAGE PROCESSING

Dr. G. Bharadwaja Kumar



Sentence Boundary Markers

- Many natural language processing (NLP) systems generally take a sentence as an input unit—part of speech (POS) tagging, chunking, and parsing, machine translation (MT), information retrieval (IR), and so forth.

- Since the errors of the SBD system propagate into the subsequent processes when they rely on accurate sentence segmentation, and the overall system performance is negatively affected.

Sentence Boundary Markers

- In some contexts period is not sentence boundary marker
- Full-stop "." has many different uses (decimal point, ellipsis, abbreviations, email and Internet addresses,..)
  - 27.5, etc., google.com, Mr.
  - The President lives in Washington D.C. He likes that place.
  - *Is K.H. Smith here?*
  - I *bought the apples, pears, lemons, etc. Did you eat them?*

Some complex Examples for SBD

(1) The group included Dr. J.M. Freeman and T. Boone Pickens Jr.

(2) "This issue crosses party lines and crosses philosophical lines!" said Rep. John Rowland (R., Conn.).

(5a) It was due Friday by 5 p.m. Saturday would be too late.

(5b) She has an appointment at 5 p.m. Saturday to get her car fixed.

- Punctuation can be present in an embedded quotation
  - Now the burning question is “How can we identify the right hyper-plane?”.

- “!” and “?” are less ambiguous but may appear in proper names (*Yahoo!*) or may be repeated to stress their meaning (*go out!!!*)

The American linguists Bernard Bloch and George L. Trager formulated the following definition: “A language is a system of arbitrary vocal symbols by means of which a social group cooperates.”

To avoid this circularity, we approximate each word’s part of speech in one of two ways: (1) by *the prior probabilities* of all parts of speech for that word, or (2) by a *binary value* for each possible part of speech for that word.

It is a reality that pluralism emerges from the very nature of our country; it is a choice made inevitable by India’s geography, re-affirmed by its history and reflected in its ethnography.

I had a ton of homework for math last night; I stayed up late to finish it all.

Often people ask questions such as "why not have an Operating System run in Tamil or Bengali?", just as Arabic or Japanese Windows.

What is a sentence here?

A few expected insights from the graph are :

1. Males in our population have a higher average height.
2. Females in our population have longer scalp hairs.

Consider the following statements:
1. Deputy Speaker and Speaker may resign by writing to each other
2. Attorney General and Solicitor General may resign by writing to each other

Which among the above statements is / are correct?

Sentence Segments

Language is a word that may be used by extension:

- The language of a community or country.
- The ability of speech.
- Formal language in mathematics, logic and computing.
- Sign language for deaf people (people who cannot hear).
- A type of school subject.

- In the Wall Street Journal (WSJ), about 42% of the periods denote abbreviations and decimal points while the corresponding percentage for Brown corpus is only 11%.

- https://www.hindawi.com/journals/tswj/2014/196574/

- The colon “:,” semicolon “;,” and comma “,” can either be a separator of grammatical subsentences or a delimiter of sentences.

- About 19% and 14% of colons are being used as the boundary delimiters of sentences in the WSJ and Brown corpus, respectively.

The list of abbreviations can be built from the training data

- An abbreviation is a token that contains a “.” that is not a sentence boundary in the training corpus.

Examples: Lib. For Library / abbr. for abbreviation / approx. for approximate

http://abbreviations.yourdictionary.com/

Character Features. We first consider the capitalization of the initial character of the word, including the immediately preceding word $f_1(w_{i-1})$ and the following word $f_2(w_{i+1})$, based on the following feature function:

$$f_{1,2}(w_i) = \begin{cases} 1 & c_1 \in \mathscr{C},\ w_i = \{c_1, c_2, \ldots, c_n\} \\ 0 & \text{otherwise,} \end{cases}$$

Word Features. Under the observation that abbreviations are generally the major source of ambiguities in the determination of sentence boundaries. They are usually short, uppercased, and tightly collocated with internal periods (i.e., acronyms) or a final period (i.e., honorific abbreviations, location name abbreviations, month and measure unit abbreviations, corporate designators, etc.) In addition to the capitalized features, we also care about the upper-/lowercasing of both the neighboring words, $f_3(w_{i-1})$ and $f_4(w_{i+1})$, defined as:

$$f_{3,4}(w_i) = \begin{cases} 1 & \forall c_i \in \mathscr{C},\ w_i = \{c_1, c_2, \ldots, c_n\} \\ 0 & \text{otherwise.} \end{cases}$$

Features five and six are considered as the length of previous and next words, that is, $f_5(w_{i-1})$ and $f_6(w_{i+1})$, respectively, and are given by

$$f_{5,6}(w_i) = |w_i|.$$

Thus, the features that used to capture this information include $f_7(w_i, w_{i+1})$, for detecting the collocation of colon with dash, period, and semicolon:

$$f_7(w_i, w_{i+1}) = \begin{cases} 1 & w_i = \text{colon}, \ w_{i+1} \in \mathscr{P}_{\text{dps}} \\ 0 & \text{otherwise}, \end{cases}$$

where $\mathscr{P}_{\text{dps}} = \{\text{dash}, \text{period}, \text{semi\_colon}\}$.

The checking of dollar sign $f_8(w_i)$ and number $f_9(w_i)$ is given by

$$f_8(w_i) = \begin{cases} 1 & w_i = \text{"\$"} \\ 0 & \text{otherwise,} \end{cases}$$

$$f_9(w_i) = \begin{cases} 1 & w_i \in \mathcal{N} \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{N}$ is numeric literals.

$f_{10}(w_i, w_{i+1})$ describes the expression of potential punctuations followed by either dash or left quotation mark:

$$f_{10}(w_i, w_{i+1}) = \begin{cases} 1 & w_i \in \mathscr{P}^*,\ w_{i+1} \in \mathscr{P}_{\text{dq}} \\ 0 & \text{otherwise.} \end{cases}$$

$\mathscr{P}^*$ is the potential punctuations that signal the boundaries of sentences, and $\mathscr{P}_{\text{dq}} = \{\text{dash}, \text{left\_quote}\}$.

$f_{11}(w_i, w_{i+1})$, on the other hand, denotes the expression that excludes left quote which immediately follows the boundary terminal $\mathscr{P}^*$ and is defined as

$$f_{11}(w_i, w_{i+1}) = \begin{cases} 1 & w_i \in \mathscr{P}^*,\ w_{i+1} \neq \mathscr{P}_q \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathscr{P}_q$ = left_quote.

Entropy & Information

- Entropy is simply the average(expected) amount of the information from the event.

Note that, to estimate entropy, the statistical properties of the source must be known, i.e. one must know what values *s* can take and how likely *(p(s))* they are.

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

Entropy in Information Theory

- Entropy is a measure of *information content:* the number of bits *actually* required to store data.
- Entropy is sometimes called a measure of *surprise*
  - A highly predictable sequence contains little actual information
    - Example: 1101101101101101101101011011 (what's next?)
  - A completely unpredictable sequence of n bits contains n bits of information
    - Example: 0100000111011001101001000O (what's next?)

Entropy in Information Theory

- As the number of possible outcomes for a random variable increases, entropy increases.

Entropy of flipping a fair coin:

$S = -(\frac{1}{2} \cdot \log_2(\frac{1}{2}) + \frac{1}{2} \cdot \log_2(\frac{1}{2})) = -2 \cdot \frac{1}{2} \cdot -1 = 1$

Entropy of rolling a die:

$S = -6 \cdot 1/6 \cdot \log_2(1/6) = -1 \cdot \log_2(1/6) = \log_2(6) = 2.585$

Principle of Maximum Entropy

- When estimating the probability distribution, you should select that distribution which leaves you the largest remaining uncertainty (i.e., the maximum entropy) consistent with your constraints.

- Maximum entropy ($H = log\ N$) is reached when all symbols are equiprobable, i.e., $p_i = 1/N$.

The Maximum Entropy Model considers only specific evidence of sentence boundaries in the text.

This evidence represents prior linguistic knowledge about contextual features of text that may indicate a sentence boundary and are determined by the experimenter.

Assumptions of the Model

If we treat a paragraph/corpus as a token stream, we can consider the SBD problem to be a random process, which delimits this paragraph/corpus into sentences. Such a process will produce an output value *y*, whose domain **Y** is all of the possible boundary positions in the stream.

In this random process, the value of *Y* may be affected by some contextual information *x*, whose domain ***X*** is all the possible textual combinations in the stream.

To solve the SBD problem, we can build a stochastic model to correctly simulate this random process. Such a model is a conditional probability model - give a context stream *x* and predict the boundaries *y* – $p(y/x)$

At the first step, we collect large number of samples – $(x_1, y_1), (x_2, y_2)\ldots(x_N, y_N)$ are extracted from the training set. We can define a joint empirical distribution over x and y from these samples:

$$\tilde{p}(x, y) = \frac{1}{N} \times number\ of\ (x,y)$$

Since the model only considers the distribution of these explicitly identified features, all other features of the text are assigned a uniform distribution.

Thus, the model is "maximally" uncertain about features of text for which it does not have prior knowledge.

We can introduce a binary function *f* for this feature:

$$f(x,y)=\begin{cases}1\ if\ \ x\ is\ a\ capitalized\ \ word\ \ following\\ \ a\ \ period\ \ y,\ \ the\ \ period\ is\ a\ boundary.\\ 0\ otherwise\end{cases}$$

We may define many features for the SBD problem.

Each of them places a constraint on the model.

The expectation of the feature from the training sample must be the same as the expectation of this feature in the model:

$$\sum_{x,y} \tilde{p}(x)\, p(y \mid x)\, f(x,y) = \sum_{x,y} \tilde{p}(x,y)\, f(x,y)$$

Where $\tilde{p}(x)$ in the constraint is the empirical distribution of *x* in the training sample.

Formula: $$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Example: Suppose a bag contains two red balls and two green balls, all of which are identical except for color. One ball is drawn at random and set aside. Then a second ball is drawn at random. Given the events

A = the first ball is red
B = the second ball is red,

find $P(B|A)$.

Answer: $P(B|A) = \frac{1}{3}$. Why? After a red ball is removed, there are two green balls left and only one red ball.

Alternatively, we could use the formula above.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(\text{both are red})}{P(\text{first is red})} = \frac{1/6}{1/2} = \frac{1}{3}.$$

the best model *p** for the SBD should maximize the conditional entroy on *p(y|x)*:

$$H(p) = -\sum_{x,y} \tilde{p}(x)\, p(y \mid x) \log\, p(y \mid x)$$

$$p^* = \arg\max_{p} H(p)$$

And subject to the following constraints at the same time:

- $p(y|x) \geq 0$. For all x,y.

$$\sum_y p(y \mid x) = 1$$

For all of the selected features (constraints).

$$\sum_{x,y} \tilde{p}(x)\, p(y \mid x)\, f_i(x,y) = \sum_{x,y} \tilde{p}(x,y)\, f_i(x,y)$$

$$i \in \{1,2...n\}$$

This is a typical constrained optimization problem, and we can use Lagrange multiplier method to solve it.

The final result of this problem after solving optimization problem using Lagrange multipliers is a log-linear (exponential) model:

$$p^*(y \mid x) = Z(x) \exp\left(\sum_i \lambda_i f_i(x, y) \right)$$

where *Z(x)*, the normalizing factor

There is no analytical method to obtain the value of λ_i

in this log-linear distribution. Therefore, we choose generalize iterative scaling as our numerical approach to obtain the vector Λ*. This iterative procedure will converge to the distribution *p**.

where Λ is a vector of weights: $\{\lambda_1, \lambda_2 ... \lambda_n\}$

Method	Precision	Recall	F1	Error Rate
RuleBased	99.56%	76.95%	86.81%	16.25%
HMM	91.43%	94.46%	92.92%	10.00%
MaxEnt	99.16%	97.62%	98.38%	1.99%

Table-1 Performance comparison of three methods