2.Finding structure of Documents

2.1 Introduction

- In human language, words and sentences do not appear randomly but have structure.
- For example, combinations of words from sentences- meaningful grammatical units, such as statements, requests, and commands.
- Automatic extraction of structure of documents helps subsequent NLP tasks: for example, parsing, machine translation, and semantic role labelling use sentences as the basic processing unit.
- Sentence boundary annotation(labelling) is also important for aiding human readability of automatic speech recognition (ASR) systems.
- Task of deciding where sentences start and end given a sequence of characters(made of words and typographical cues) **sentences boundary detection.**
- **Topic segmentation** as the task of determining when a topic starts and ends in a sequence of sentences.

- The statistical classification approaches that try to find the presence of sentence and topic boundaries given human-annotated training data, for segmentation.
- These methods base their predictions on features of the input: local characteristics that give evidence toward the presence or absence of a sentence, such as a period(.), a question mark(?), an exclamation mark(!), or another type of punctuation.
- Features are the core of classification approaches and require careful design and selection in order to be successful and prevent overfitting and noise problem.
- Most statistical approaches described here are language independent, every language is a challenging in itself.
- For example, for processing of Chinese documents, the processor may need to first segment the character sequences into words, as the words usually are not separated by a space.
- Similarly, for morphological rich languages, the word structure may need to be analyzed to extract additional features.
- Such processing is usually done in a pre-processing step, where a sequence of tokens is determined.
- Tokens can be word or sub-word units, depending on the task and language.
- These algorithms are then applied on tokens.

2.1.1 Sentence Boundary Detection

- **Sentence boundary detection** (Sentence segmentation) deals with automatically segmenting a sequence of word tokens into sentence units.
- In written text in English and some other languages, the beginning of a sentence is usually marked with an uppercase letter, and the end of a sentence is explicitly marked with a period(.), a question mark(?), an exclamation mark(!), or another type of punctuation.
- In addition to their role as sentence boundary markers, capitalized initial letters are used distinguish proper nouns, periods are used in abbreviations, and numbers and punctuation marks are used inside proper names.
- The period at the end of an abbreviation can mark a sentence boundary at the same time.
- Example: I spoke with Dr. Smith. and My house is on Mountain Dr.
- In the first sentence, the abbreviation Dr. does not end a sentence, and in the second it does.
- Especially **quoted sentences** are always problematic, as the speakers may have uttered multiple sentences, and sentence boundaries inside the quotes are also marked with punctuation marks.
- An automatic method that outputs word boundaries as ending sentences according to the presence of such punctuation marks would result in cutting some sentences incorrectly.

- Ambiguous abbreviations and capitalizations are not only problem of sentence segmentation in written text.
- Spontaneously written texts, such as short message service (SMS) texts or instant messaging(IM) texts, tend to be nongrammatical and have poorly used or missing punctuation, which makes sentence segmentation even more challenging.
- Similarly, if the text input to be segmented into sentences comes from an **automatic system**, such as optical character recognition (OCR) or ASR, that aims to translate images of handwritten, type written, or printed text or spoken utterances into machine editable text, the finding of sentences boundaries must deal with the errors of those systems as well.
- On the other hand, for conversational speech or text or multiparty meetings with ungrammatical sentences and disfluencies, in most cases it is not clear where the boundaries are.
- Code switching -that is, the use of words, phrases, or sentences from multiple languages by multilingual speakers- is another problem that can affect the characteristics of sentences.
- For example, when switching to a different language, the writer can either keep the punctuation rules from the first language or resort to the code of the second language.

- Conventional rule-based sentence segmentation systems in well-formed texts rely on patterns to identify potential ends of sentences and lists of abbreviations for disambiguating them.
- For example, if the word before the boundary is a known abbreviation, such as “Mr.” or “Gov.,” the text is not segmented at that position even though some periods are exceptions.
- To improve on such a rule-based approach, sentence segmentation is stated as a classification problem.
- Given the training data where all sentence boundaries are marked, we can train a classifier to recognize them.

2.1.2 Topic Boundary Detection

- **Segmentation**(Discourse or text segmentation) is the task of automatically dividing a stream of text or speech into topically homogenous blocks.
- This is, given a sequence of(written or spoken) words, the **aim of topic segmentation** is to find the boundaries where topics change.

- Topic segmentation is an important task for various language understanding applications, such as information extraction and retrieval and text summarization.
- For example, in information retrieval, if a long documents can be segmented into shorter, topically coherent segments, then only the segment that is about the user's query could be retrieved.
- During the late1990s, the U.S defence advanced research project agency(DARPA) initiated the **topic detection and tracking program** to further the state of the art in finding and following **new topic** in a stream of broadcast news stories.
- One of the tasks in the TDT effort was segmenting a **news stream into individual stories**.

2.2 Methods

- Sentence segmentation and topic segmentation have been considered as a **boundary classification problem**.
- Given a boundary candidate(between two word tokens for sentence segmentation and between two sentences for topic segmentation), the goal is to predict whether or not the candidate is an actual boundary (sentence or topic boundary).

- Formally, let **x∈X be the vector of features** (the observation) associated with a candidate and **y ∈Y be the label** predicted for that candidate.
- The label y can be **b for boundary** and $\bar{b}$ **for nonboundary**.
- Classification problem: given a set of training examples$(x,y)_{train}$, find a function that will assign the most accurate possible label y of unseen examples $x_{unseen.}$
- Alternatively to the binary classification problem, it is possible to model boundary types using finer-grained categories.
- For segmentation in text be framed as a three-class problem: sentence boundary b^a, without an abbreviation b^a and abbreviation not as a boundary $\bar{b}^a$.
- Similarly spoken language, a three way classification can be made between non-boundaries $\bar{b}$ statements b^s, and question boundaries b^q .
- For sentence or topic segmentation, the problem is defined as finding the most probable sentence or topic boundaries.
- The natural unit of sentence segmentation is words and of topic segmentation is sentence, as we can assume that topics typically do not change in the middle of a sentences.

- The words or sentences are then grouped into categories stretches belonging to one sentences or topic- that is word or sentence boundaries are classified into sentences or topic boundaries and -non-boundaries.
- The classification can be done at each potential boundary *i* (local modelling); then, the aim is to estimate the most probable boundary type $\hat{y}_i$ for each candidate x_i

$$\hat{y} = \underset{y_i \, in \, Y}{argmax} P(y_i|x_i)$$

Here, the ^ is used to denote estimated categories, and a variable without a ^ is used to show possible categories.

- In this formulation, a category is assigned to each example in isolation; hence, decision is made locally.
- However, the consecutive types can be related to each other. For example, in broadcast news speech, two consecutive sentences boundaries that form a single word sentence are very infrequent.
- In local modelling, features can be extracted from surrounding example context of the candidate boundary to model such dependencies.

- It is also possible to see the candidate boundaries as a sequence and search for the sequence of boundary types $\hat{Y} = \hat{y_1}, \ldots, \hat{y_n}$ that have the maximum probability given the candidate examples, $X = x_1, \ldots, x_n$:

$$\hat{Y} = \underset{y}{argmax} P(Y|X)$$

- We categorize the methods into local and sequence classification.
- Another categorization of methods is done according to the type of the machine learning algorithm: **generative versus discriminative**.
- Generative sequence models estimate the **joint distribution** of the observations P(X,Y) (words, punctuation) and the labels(sentence boundary, topic boundary).
- Discriminative sequence models, however, **focus on features** that categorize the differences between the labelling of that examples.

2.2.1 Generative Sequence Classification Methods

- Most commonly used generative sequence classification method for topic and sentence is the hidden Markov model (HMM).
- The probability in equation 2.2 is rewritten as the following, using the Bayes rule:

$$\hat{Y} = \underset{y}{argmax} P(Y|X) \quad 2.1$$

$$\hat{Y} = \underset{y}{argmax} P(Y|X) = \underset{y}{argmax} (P(X|Y)P(Y)/P(X)) = \underset{y}{argmax} (P(X|Y)P(Y) \quad 2.2$$

Here $\hat{Y}$= Predicted class(boundary) label

Y = (y_1,y_2,….y_k)= Set of class(boundary) labels

X = (x_1,x_2,….x_n)= set of feature vectors

P(Y|X) = the probability of given the X (feature vectors), what is the probability of X belongs to the class(boundary) label.

P(x) = Probability of word sequence

P(Y) = Probability of the class(boundary)

$$\hat{Y} = \underset{Y}{\operatorname{argmax}} P(Y|X) = \underset{Y}{\operatorname{argmax}} \frac{P(X|Y)P(Y)}{P(X)} = \underset{Y}{\operatorname{argmax}} P(X|Y)P(Y) \qquad (2.3)$$

- P(X) in the denominator is dropped because it is fixed for different Y and hence does not change the argument of max.
- P(X|Y) and P(Y) can be estimated as

$$P(X|Y) = \prod_{i=1}^{n} P(\mathbf{x}_i|y_1, \ldots, y_i) \tag{2.4}$$

and

$$P(Y) = \prod_{i=1}^{n} P(y_i|y_1, \ldots y_{i-1}) \tag{2.5}$$

2.2.2Discriminative Local Classification Methods

- Discriminative classifiers aim to model **$P(y_i | x_i)$ equation 2.1 directly**.
- The most important distinction is that whereas c**lass densities P(x|y)** are model assumptions **in generative approaches**, such as naïve Bayes, in discriminative methods, discriminant functions of the feature space define the model.
- A number of discriminative classification approaches, such as support vector machines, boosting, maximum entropy, and regression. Are based on very different machine learning algorithms.
- While discriminative approaches have been shown to outperform generative methods in many speech and language processing tasks.
- For **sentence segmentation, supervised learning methods have primarily been applied to newspaper articles**.
- Stamatatos, Fakotakis and Kokkinakis used **transformation based learning** (TBL) to infer rules for **finding sentence boundaries**.

- Many classifiers have been tried for the task: regression trees, neural networks, classification trees, maximum entropy classifiers, support vector machines, and naïve Bayes classifiers.
- The most Text tiling method Hearst for topic segmentation uses a l**exical cohesion metric** in a word vector space as an indicator of topic similarity.
- Figure depicts a typical graph of similarity with respect to **consecutive segmentation units**.

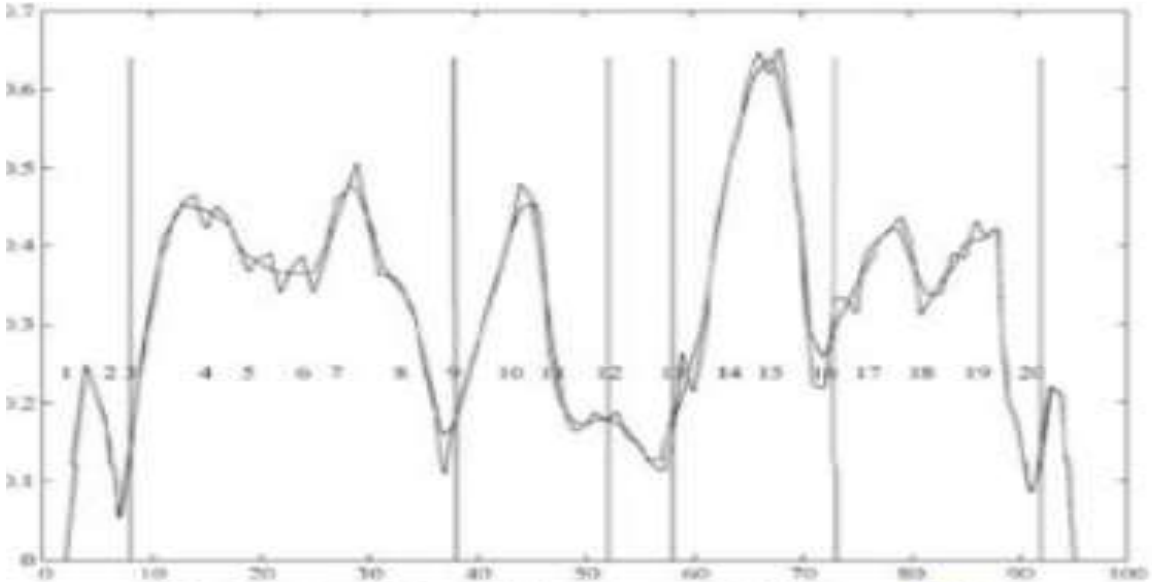

Figure 2–4. Text Tiling example (from [22])

Originally, two methods for computing the similarity scores were

- The document is chopped when the similarity is below some threshold.
- Originally, **two methods** for computing the similarity scores were proposed: **block comparison** and **vocabulary introduction**.

- The first, **block comparison, compares adjacent blocks of text to see how similar** they are according to how many words the adjacent blocks have in common.
- **Given two blocks, b_1 and b_2, each having k tokens** (sentences or paragraphs), the **similarity (or topical cohesion) score** is computed by the **formula**:

$$\frac{\sum_t w_{t,b_1} \cdot w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}}$$

- **Where $w_{t,b}$ is the weight assigned to term t in block b**.
- The weights can be binary or may be computed using other information retrieval- metrics such as term frequency.
- The second, the **vocabulary introduction method**, **assigns a score to a token-sequence gap** on the basis of **how many new words are seen in the interval in which it is the midpoint**.

- Similar to the **block comparison formulation, given two consecutive blocks b_1 and b_2, of equal number of words w**, the topical cohesion score is computed with the following formula:

$$\frac{NumNewTerms(b_1) + NumNewTerms(b_2)}{2 \times w}$$

- Where **NumNewTerms(b) returns the number of terms in block b seen the first time in text**.

2.2.3Discriminative Sequence Classification Methods

- In segmentation tasks, the sentence or topic decision for a given example(word, sentence, paragraph) highly depends on the decision for the examples in its vicinity.
- Discriminative sequence classification methods are in general extensions of local discriminative models with additional decoding stages that find the best assignment of labels by looking at neighbouring decisions to label.
- Conditional random fields(CRFs) are extension of maximum entropy, SVM struct is an extension of SVM, and maximum margin Markov networks(M^3N) are extensions of HMM.
- CRFs are a class of log-linear models for labelling structures.

- Contrary to local classifiers that predict sentences or topic boundaries independently, CRFs can oversee the whole sequence of boundary hypotheses to make their decisions.

Complexity of the Approaches

- The approaches described here have **advantages** and **disadvantages**.
- In a **given context** and under a set of observation features, one **approach may be better than other**.
- These approaches can be rated in terms of **complexity** (time and memory) of **their training** and **prediction algorithms** and in terms of their **performance on real-world datasets**.
- In terms of complexity, **training of discriminative approaches** is **more complex than training of generative ones** because they require multiple passes over the training data to adjust for feature weights.
- However, generative models such as HELMs can handle **multiple orders of magnitude larger training sets** and benefits, for instance, from decades of news wire transcripts.
- On the other hand, they work with **only a few features** (only words for HELM) and do not cope well with unseen events.