1. Topic Detection Algorithm

In the TDT3 formal evaluation, our group attends topic detection task. We focus on Mandarin text. Because the original data in TDT3 is in GB codes (simplified Chinese characters), the first step is to convert the data into Big-5 codes (traditional Chinese characters). After dividing the converted files into pieces of news stories, we employ a Chinese named entity extraction system used in MUC7 to identify people names, organization names, location names, and some other named entities like date/time expressions and monetary and percentage expressions.

At most 50 terms are selected from each news story to represent its content. In the version used in TDT3 formal run, only named entities are used. The weight of each term is measured as follows:

\[
\text{term weight} = \frac{\text{term frequency}}{\text{number of topics detected up to now containing this term}}
\]

A news story is represented as a term vector \( V_d \). Similarly, a topic is represented as a term vector \( V_t \). The similarity measure is defined as follows:

\[
\text{similarity measure } S_{td} = \frac{V_t \cdot V_d}{|V_t||V_d|}
\]

Two thresholds, low threshold (\( TH_l \)) and high threshold (\( TH_h \)), are specified. Their interpretations are given below:

- if \( S_{td} < TH_l \) then news story \( d \) is irrelevant to topic \( t \);
- if \( S_{td} \geq TH_h \) then news story \( d \) is relevant to topic \( t \);
- if \( TH_l \leq S_{td} < TH_h \) then the relationship between \( d \) and \( t \) is undecidable.

Initially, the first news story \( d_1 \) is assigned to topic \( t_1 \). Assume there have been \( k \) topics when a new document \( d_i \) is considered. That is, topics \( t_1, t_2, \ldots, t_k \) (\( k \leq i \)) have been detected. The news story \( d_i \) may belong to one of \( k \) topics, or it may form a new topic \( t_{k+1} \). That is determined by the similarity measure defined before. If there exists a topic \( t_k \) such that \( d_i \) is relevant to \( t_k \), then we say the news story touches on an old event. We select a topic, say \( t_k \), with the highest similarity with \( d_i \), and insert \( d_i \) into \( t_k \). The term vector \( V_{t_k} \) is changed accordingly. In contrast to this case, if \( d_i \) is irrelevant to all the topics, then we say the news story deals with a new event.
If the similarity is not above the high threshold or below the low threshold, then it is not decidable. In our algorithm, we should consider the next 10 news stories further. Some of these news stories may belong to topics \( t_1, \ldots, t_k \). Recall that the term vectors for the corresponding topics may be changed when such news stories are inserted. Under such a situation, the similarity measure between the changed topic and the undecidable news story may above the high threshold or below the low threshold. In other words, relevance or irrelevance may be decided after the next 10 news stories are read. If it is still undecidable, we change the decision procedure in the following way. Here, \( TH_m \) is equal to \((TH_l + TH_h)/2\).

\[
\text{if } TH_l \leq \text{new } S_{td} < TH_m \text{ then } d \text{ is irrelevant to } t.
\]

\[
\text{if } T_m \leq \text{new } S_{td} < TH_h \text{ then } d \text{ is relevant to } t.
\]

The conference value \( C_f \) is defined as follows:

1. document \( d \) is relevant to topic \( t \)
   \[
   C_f = (S_{td} - TH_m)/(1 - TH_m)
   \]
2. document \( d \) is irrelevant to topic \( t \)
   \[
   C_f = (TH_m - S_{td})/TH_m
   \]

2. Experimental Environments

The test material is selected from TDT3 corpus. There are 9130 pieces of news stories. The topic detection algorithm was run on a Pentium III-500 with 128M memory. The performance is listed below:

(1) SR=nwt+bnasr TE=man,nat boundary DEF=10
    story-weighted: P(miss)=0.8255, P(fa)=0.0004, (Cdet)norm=0.8277
    story-weighted: P(miss)=0.7282, P(fa)=0.0004, (Cdet)norm=0.7303

(2) SR=nwt+bnman TE=man,nat boundary DEF=10
    story-weighted: P(miss)=0.8250, P(fa)=0.0005, (Cdet)norm=0.8273
    story-weighted: P(miss)=0.7268, P(fa)=0.0005, (Cdet)norm=0.7291

Two sets of thresholds are adopted. For NTU1, \( TH_l=01 \) and \( TH_h=0.5 \). For NTU2, \( TH_l=02 \) and \( TH_h=0.4 \).

3. Discussions

In the experiments, the thresholds are not tuned. Our threshold is too high, so that too many new topics are produced. In other words, Miss is high and False Alarm is almost zero. Compared NTU1 and NTU2, to enlarge threshold period seems not to have much influence. Now we are studying the setting of \( TH_l \) and \( TH_h \), and the extension of our methods to English and multilingual corpora. Besides, verbs and other nouns are considered as feature candidates in this study.