CRUISE at the NTCIR-10 Mission Impossible Task

Tom A. Cruise
Disavowed, U.S.A
toma@cruise.com

Tom B. Cruise
Disavowed, U.S.A
tomb@cruise.com

Tom C. Cruise
Disavowed, U.S.A
tomc@cruise.com

Tom D. Cruise
Disavowed, U.S.A
tomd@cruise.com

Tom E. Cruise
Disavowed, U.S.A
tome@cruise.com

Tom F. Cruise
Disavowed, U.S.A
The Last Samurai
Corporation, Japan
tomf@cruise.com

ABSTRACT

The CRUISE team participated in the Climb the Dubai Tower (CDT) subtask of the NTCIR-10 Mission Impossible Task. This minority report describes our approach to solving the CDT problem and discusses the official results.

Team Name

CRUISE

Subtasks

Climb the Dubai Tower (Chinese, English, Japanese)

Keywords

eyes wide shut, top gun

1. INTRODUCTION

The CRUISE team participated in the Climb the Dubai Tower (CDT) subtask of the NTCIR-10 Mission Impossible Task [1]. This minority report describes our approach to solving the CDT problem and discusses the official results.

2. METHOD

2.1 RUN1

2.2 RUN2

We adopt a method that using a word co-occurence network. This method has following 3 parts:

1. Make Network
2. Extract Subnetwork
3. Output Rank Result

2.2.1 Make Network

In this part, we make a word co-occurence network from pairs of post tweets and reply tweet. As well as in the previous section, the data has 0000000 tweets. First, we analyze morphenom against all tweet pairs by MeCab. Then we extract pair of sets of noun words from each pair as (W_p, W_r). Wp or Wr contains a ordered word set which means faster word appear faster in the original tweet. And set V as an union of all pairs.

1. $W_p = (\omega_{p1}, \omega_{p2}, ...\omega_{pi}...\omega_{pn})$
2. $W_r = (\omega_{r1}, \omega_{r2}, ...\omega_{ri}...\omega_{rn})$
3. $P = W|W \in \bigcup W_p \wedge W \in \bigcup W_r$
4. $V = \bigcup(W_p, W_r)$

Using V as nodes, we make a directed graph. We make a path from the first word in some tweet to the next word in the tweet and from each word in W_p to all words in W_r.

$$E = \{(\omega_1, \omega_2)|\omega_1 \in W_p \wedge \omega_2 \in W_r\} \cup \{(\omega_{pi}, \omega_{p(i+1)})|0 < i < n-1\} \cup \{(\omega_r$$

At last a word co-occurence network G is defined like below:

$$G := (V, E)$$

This word network shows what word likely appear next to some word. In other words, it shows that how wrods close to each other. So, it represents word distribution which contains topic distribution.

2.2.2 Extract Subnetwork

Test tweet data is same as the data used in RUN1. As well as previous part, we extract set of noun words from each test tweet as $W_t = (\omega_{t1}, \omega_{t2}, ...\omega_{ti}...\omega_{tn})$. In this part, we extract subnetwork G'_i which contains ω_{ti} from G. If wti exists in V, pick up all nodes which is next to wti.

$$V'_i = \omega_{ti} \cup \{\omega_e|(\omega_{ti}, \omega_e) \in E\}$$

$$E'_i = \{(\omega_{ti}, \omega_e)|(\omega_{ti}, \omega_e) \in E\}$$

Finaly, we get a subnetwork G' like below:

$$G'_i = (V'_i, E'_i)$$

$G' = \bigcup G'_i$ represents potential topics of expected replies.

2.2.3 Output Rank Result

In this part, we get results for each W_t using subnetworks we get previous part. We rank each possible reply using tf-idf like score and pagerank.

First, we calculate pagerank for each word in W_t. With $E(\omega_k)$, which is the number of edges that has ω_k as the start node, a pagerank is calculated like below:

$$PR(\omega_{ti}) = (1-d)/N + \sum_{\omega_k \in V_i'} d(PR(\omega_k)/E(\omega_k))$$

Then, we calculate final score with tf-idf like score. P' is a set of tweets which concludes ω_{ti}.

$$P_i' = \{W_i | W_i \in P \land \omega_{ti} \in W_i\}$$

And, we define df as the number of tweets in P', N as the number of tweets in P, l as the number of word in W_i, and tc as how many ω_{ti} appear in W_i.

1. $df = n(P')$
2. $N = n(N)$
3. $l = n(W_i)$
4. $tc = n(\{\omega | \omega = \omega_{ti} \land \omega \in W_i\})$

$$Score_{TWEET} = tfidf(W_t)$$

2.2.4 Two types of system

3. ADDITIONAL AUTHORS

Additional authors: Tom G. Cruise (Disavowed, U.S.A. email: tomg@cruise.com) and Tom H. Cruise (Disavowed, U.S.A. email: tomh@cruise.com).

4. REFERENCES

[1] T. I. Cruise and T. J. Cruise. Overview of the NTCIR-10 mission impossible task. In Proceedings of NTCIR-10, pages 1–100, 2013.