11 Introduction

- Cnt’d

11.1 Introduction

- SMP: aimed to discover cross-lingual links of actual Wikipedia, in other words it targets Wikipedia Link Discovery (CLLD).
  - REF: aimed to discover as much meaningful cross-lingual links as possible automatically.

12-1 Extracting Candidate Anchor Text

- SMP: get the candidate anchor texts as anchor text of Web page (written in source language) in the Internet retrieved by title of each topic.
- REF: makes the candidate anchor texts of every string (grams) in the fixed range length from each topic.

12-2 Filtering Candidate Anchor Text

- For both type of runs, the candidate anchor texts are filtered by the following two steps.
  - [step 1]: They should match a title of a pages in source language corpus.
  - [step 2]: Sub-strings which are included in a longer string is abandoned, even if they match title.
- [step 2]: There are links from the page step 1 to pages of target language corpus.

12-3 Getting Cross-lingual Link

- In step 1, links in matched page are extracted.
- In step 2, by following ‘other language’ part of the searched page we get links to the pages of target language corpus as cross-lingual links.

13-1 Experimental Results

- Environment of experiment.
  - CPU: Intel Corei5-3570G @ 3.6GHz 4C/4T
  - Memory: 4GB
  - HDD: SATA300 500GB 7200 RPM 16MB buffer
  - OS: Fedora 19 64-bit
  - Programming Language: C, Perl, R 3.2

- Submitted run REF (OKSAT-(C2JKE2E2JK)-A2F-01-REF) is aimed to discover as much meaningful cross-lingual links as possible automatically.
- So, we are interested in the number of links and anchor texts.

14-1 Topic by Topic

- Figure 3(a), (b), and (c) show the number of links of SMP and REF in C2JKE2 subtask.
- In this figure, the topic-id is in horizontal, and SL (the number of links of before ‘Get links to target language pages’ in Figure 1) and TL (that of after) are in vertical.
- It turns out that the number of links in REF is more than that of SMP in any topic.
- From SL (links in source language) to TL (links in target language), we know how these links are filtered.

14-2 Subtask by Subtask

- Figure 5 shows the total number of links of qrels assessed by organizer, that is A2BWikiManualResultSet-(C2JKE2E2JK).xml (MAN in this figure) and REF over topics for each subtasks.