DCU at VideoCLEF 2009

Ágnes Gyarmati and Gareth J. F. Jones
Centre for Digital Video Processing
Dublin City University, Dublin 9, Ireland
{agyarmati|gjones}@computing.dcu.ie

Abstract

DCU participated in the VideoCLEF 2009 Linking Task. Our approach was based on identifying relevant related content using the Lemur information retrieval toolkit. We implemented two distinctive variants of our approach. One version performs the search in the Dutch Wikipedia with the exact words (either stemmed or not) of the search query extracted from the ASR transcription, and returns the corresponding links pointing to the English Wikipedia. The other variant first performs an automatic machine translation of the Dutch query into English, and then the translated query is used to search the English Wikipedia directly. Among our four runs, we achieved the best results with the first approach, when the base of retrieval was the stemmed and stopped Dutch Wikipedia. Unfortunately for us, there is no one-to-one relation between the pages of the Dutch and the English Wikipedias, hence some hits from the Dutch Wikipedia have been lost as results due to lack of equivalent English article. In extreme cases, our system might return no output at all if none of the hits for a given anchor are linked to a page in the English Wikipedia. Although we included a preprocessing phase before indexing the article collections, some unuseful, but frequently occurring types of page escaped and had a significant negative impact of our second basic approach implemented in Run 3.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing—*Speech recognition and synthesis*

General Terms

Information Retrieval, Automatic Speech Recognition

Keywords

information retrieval, automatic speech recognition, cross-language description linking

1 Introduction

The VideoCLEF Linking Task involves locating content related to sections of an automated speech recognition (ASR) transcription cross-lingually. Elements of a Dutch ASR transcription are to be linked to related pages in an English Wikipedia collection. We submitted four runs for the VideoCLEF 2009 Linking Task [1], by implementing two different approaches to solve the task. Because of the difference between the source language (Dutch) and the target language (English),

a switch between the languages at some point in the system is inevitable. The two approaches differ in defining the switching method.

One approach performs the search in the Dutch Wikipedia with the exact words (either stemmed or not) of the search query extracted from the ASR transcription, and returns the corresponding links pointing to the English Wikipedia. The other variant first performs an automatic machine translation of the Dutch query into English, and then the translated query is used to search the English Wikipedia directly.

2 System Description

Wikipedia dumps are created and published regularly, for our experiments we used the dump dated May 30th 2009 for the English, and the dump dated May 31st 2009 for the Dutch language Wikipedia collection. In a simple prepocessing phase, we eliminated some information we did not consider as relevant in the point of the task, e.g. information about users, comments, links to other languages we did not need. For indexing and retrieving, we used the *Indri* model of the open source Lemur Toolkit [2], unaltered, off-the-shelf. English texts were stemmed by Lemur's built-in stemmer, while Dutch texts were stemmed by using Oleander's implementation [5] of `Snowball`'s Dutch stemmer algorithm [6]. We used stopword lists provided by `Snowball` for both languages.

Queries were formed based on the sequences of words extracted from the ASR transcripts for each of the anchors defined by the task. As transcript files contain timing information for each word, and anchors were defined by their starting and end point, respectively, our system searched for the given starting point of an anchor in the transcript file, and took all the words consecutively that fell in the time period until the given end point. First, these sequences were used directly as queries for retrieval from the Dutch collection, and the Dutch Wikipedia's own links pointing to the corresponding articles of the English version were returned for each anchor point and each retrieved article, as the solution for the task. The other option was to translate the words to English first, and searching in the English Wikipedia using the translation as query. Translations were performed automatically using the query translation component developed for the `Multimatch` project [3]. This translation tool combines the WorldLingo machine translation engine augmented with a bilingual dictionary from the cultural heritage domain automatically extracted from the multilingual Wikipedia.

3 Run Configurations

Here we describe the four runs we submitted to the Linking Task. The most prominent feature of each run is the choice of the collection for retrieval, i.e. whether it was the Dutch or the English Wikipedia.

1. **Dutch** As `Lemur` does not have Dutch-specific built-in tools (stemmer), we indexed the Dutch wikipedia as it was, without stemming or stopping. Retrieval was then performed from the Dutch collection, returning the relevant links found there.

2. **Dutch stemmed** The steps of retrieval are identical to that of Run 1, the only difference lies in the processing of the collection (and of queries), text is stemmed and stopped.

3. **English** This run represents the second approach, with the query translated first and retrieval then performed in the English collection. Text was stemmed and stopped.

4. **Dutch with blind relevance feedback** This run is almost identical to Run 1, with a difference in parameter setting for `Lemur` to perform blind relevance feedback. `Lemur/Indri` uses a relevance model, for details see [4]. The first ten retrieved documents were assumed relevant and queries were expanded by five terms.

4 Results

In this section we present the results obtained by our various runs. The Linking Task was assessed by the organisers as a known item task. The top most relevant link for each anchor is called a *primary* link, and all other relevant links defined additionally by the assessors are called *secondary* links [1].

Table 1 lists Recall and Mean Reciprocal Rank (MRR) for primary links, Table 2 shows only MRR values for secondary links as Recall cannot be counted due to the lack of an exhaustive list of secondary links.

Run	Recall	MRR
Run 1	44/165	0.18202
Run 2	44/165	0.18202
Run 3	13/165	0.05606
Run 4	38/165	0.14434

Table 1: Scores for Primary Links

Runs 1 and 2 achieved the highest scores. Although they do yield slightly different output, the decision on whether to stem and stop text does not alter the results statistically, in the matter of primary links, while stemming and stopping (Run 2) improved results in finding secondary links. Run 4 used blind relevance feedback to expand the queries, setting the optimal parameters for these process would require further experimentation, and other expansion methods than Indri's relevance model can be used and results compared.

The main problem of this approach (that is, addressing the Dutch collection) lies in the differences between the English and the Dutch versions of Wikipedia. Although the English site is approximately ten times larger than its Dutch counterpart (considering the number of articles), there are articles that have no equivalent page in the other language, due to different structuring on the other side, or cultural differences, for example. System 1, 2 and 4 might (and in fact did) come up with relevant links at some point which were lost when checking for direct links provided on the Dutch page pointing to the English page. A weak point of out system in this approach is that some hits from the Dutch Wikipedia might get lost as results due to the lack of an equivalent English article. In extreme case, our system might return no output at all if none of the hits for a given anchor are linked to any page in the English Wikipedia.

Run	MRR
Run 1	0.26773
Run 2	0.27475
Run 3	0.08990
Run 4	0.18960

Table 2: Scores for Related Links

Run 3, which involved the implementation of a different approach, performed significantly worse. This might originate due to two aspects of the switch to the English collection. First, the query text was translated automatically from Dutch to English, which in itself carries a certain risk of losing information due to misinterpreting words or expressions or ignoring words unrecognised by the translation tool. While MultiMatch translation tool has a vocabulary expanded to include many concepts from the domain of cultural heritage, there were many specialist concepts in the ASR transcription which are not included in its translation vocabulary. Approximately 3.5% of the words were left in Dutch unchanged by the translator (names not included) which might be considered as insignificant, but some of them turned out to be key words (e.g. *rariteitenkabinet* 'cabinet of curiosities', which was in fact retrieved by the system for Run 1 and 2 (although ranked

lower than desired)).

The other main problem we encountered at Run 2 lay in the size of the English Wikipedia and our insufficient experience concerning its structure. The downloadable dump includes a large number of pages that look like useful articles, but are in fact not: *used to be* or *not meant* to be articles at all (discussion pages, articles for deletion). This phenomenon missed our attention during the development phase, but had a high impact on our results, as about 18.5 % of the links given by Run 3 as solution were proven as invalid. Cleaning up the data more carefully gives a fairer opportunity to system 3.

Results are also (mostly negatively) affected by the quality of ASR transcripts. As transcripts were provided with the task, we used them as they were in each of the runs, it was not the transcripts but the steps followed that distinguished the four runs. For more discussion on the relation of the transcripts to the Linking Task, see [1].

5 Conclusions

In this paper we outlined details of our submissions to the Linking Task at VideoCLEF 2009. We described how data were processed and how the system, or systems, worked, and presented the results each run scored.

On the one hand, there is still room for improvement from our side, e.g. by finding better parameter settings and/or other expansion methods, by preprocessing data better by eliminating more unuseful information than we had previously done.

On the other hand, improvement in independent factors might help our system to achieve better results as well. To mention one possibility, wikipedia.org grows larger and larger, in every language, editors might add the links we were missing.

Acknowledgements

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008. We are grateful to Eamonn Newman for assistance with the MultiMatch translation tool.

References

[1] Larson, Martha, Newman, Eamonn and Jones, Gareth J. F. Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment. In Borri, Francesca and Nardi, Alessandro and Peters, Carol (eds.) *Working Notes of CLEF 2009*

[2] The Lemur Toolkit. http://www.lemurproject.org/

[3] Jones, Gareth J. F., Fantino, Fabio, Newman, Eamonn and Zhang, Ying. Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia, In *Proceedings of the 2nd International Workshop on Cross Lingual Information Access - Addressing the Information Need of Multilingual Societies (CLIA-2008)*, Hyderabad, India, pp34-41, 2008.

[4] Metzler, Don. *Indri Retrieval Model Overview.*
http://ciir.cs.umass.edu/ metzler/indriretmodel.html

[5] Oleander Stemming Library. http://sourceforge.net/projects/porterstemmers/

[6] Snowball. http://snowball.tartarus.org/