When to Cross Over? Cross-Language Linking Using Wikipedia for 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. We describe Dublin City University (DCU)'s participation in the VideoCLEF 2009 Linking Task. Two approaches were implemented using the Lemur information retrieval toolkit. Both approaches first extracted a search query from the transcriptions of the Dutch TV broadcasts. One method first performed search on a Dutch Wikipedia archive, then followed links to corresponding pages in the English Wikipedia. The other method first translated the extracted query using machine translation and then searched the English Wikipedia collection directly. We found that using the original Dutch transcription query for searching the Dutch Wikipedia yielded better results.

1 Introduction

The VideoCLEF Linking Task involved locating content related to sections of an automated speech recognition (ASR) transcription cross-lingually. Elements of a Dutch ASR transcription were to be linked to related pages in an English Wikipedia collection [1]. We submitted four runs 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 is required at some point in the system. Our two approaches differed in the switching method.

One approach performed the search in a Dutch Wikipedia archive with the exact words (either stemmed or not) and then returned the corresponding links pointing to the English Wikipedia pages. The other one first performed an automatic machine translation of the Dutch query into English, the translated query was then used to search the English Wikipedia archive directly.

2 System Description

For our experiments we used the Wikipedia dump dated May 30th 2009 for the English archive, and the dump dated May 31st 2009 for the Dutch Wikipedia collection. In a simple preprocessing phase, we eliminated some information irrelevant to 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]. English texts were stemmed using Lemur's built-in stemmer, while Dutch texts were stemmed 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 sequences of words extracted from the ASR transcripts using the word timing information in the transcript file. For each of the anchors defined by the task, the transcript was searched from the anchor starting point until the given end point, and the word sequence between these boundaries extracted as the query. These sequences were used directly as queries for retrieval from the Dutch collection. The Dutch Wikipedia's links pointing to the corresponding articles of the English version were returned as the solution for each anchor point in the transcript. For the other approach queries were translated automatically from Dutch to English 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. The translated query was used to search the English Wikipedia archive.

3 Run Configurations

Here we describe the four runs we submitted to the Linking Task, plus an additional one performed subsequently.

1. **Dutch** The Dutch Wikipedia was indexed without stemming or stopping. Retrieval was performed on the Dutch collection, returning the relevant links from the English collection.
2. **Dutch stemmed** Identical to Run 1, except that the Dutch Wikipedia text is stemmed and stopped as described in Sect. 2.
3. **English** This run represents the second approach with stop word removal and stemming applied to the English documents and queries. The translated query was applied to the indexed English Wikipedia.
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 query expansion, for details see [4]. The first 10 retrieved documents were assumed relevant and queries were expanded by 5 terms.
5. **English** (referred to as 3′) This is an amended version of Run 3, with the difference of an improved preprocessing phase applied to the English Wikipedia, disregarding irrelevant pages as described in Sect. 4.

4 Results

The Linking Task was assessed by the organisers as a known item task. The top ranked relevant link for each anchor is referred to as a *primary* link, and all other relevant links identified by the assessors as *secondary* links [1].

Table 1. Scores for Related Links

Run	Recall (prim)	MRR (prim)	MRR (sec)
Run 1	0.267	0.182	0.268
Run 2	0.267	0.182	0.275
Run 3	0.079	0.056	0.090
Run 4	0.230	0.144	0.190
Run 3′	0.230	0.171	–

Table 1 shows Recall and Mean Reciprocal Rank (MRR) for primary links, and MRR values for secondary links. Recall cannot be calculated for secondary links due to the lack of an exhaustive identification of secondary links. Table 1 also includes Run 3′ evaluated automatically using the same set of primary links as in the official evaluation. Secondary links have been omitted as we could not provide the required additional manual case-by-case evaluation by the assessors.

Runs 1 and 2 achieved the highest scores. Although they do yield slightly different output, the decision of whether to stem and stop text does not alter the results statistically, in the matter of primary links, while stemming and stopping (Run 2) did improve results a little in finding secondary links. Run 4 using blind relevance feedback to expand the queries was not effective here. Setting the optimal parameters for this process would require further experimentation, and either this or alternative expansion methods may produce better results.

The main problem of retrieving from the Dutch collection lies in the differences between the English and the Dutch versions of Wikipedia. Although the English site contains a significantly larger number of articles, there are articles that have no equivalent pages cross-lingually, due to different structuring or cultural differences. Systems 1, 2 and 4 might (and in fact did) come up with relevant links at some points which were lost when looking for a direct link to an English page. Thus a weak point of our approach is that some hits from the Dutch Wikipedia might get lost in the English output due to the lack of an equivalent English article. In the 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 3 performed significantly worse. This might be 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 risk of translation errors due to misinterpretation of the query or weaknesses in the translation dictionaries. While the MultiMatch translation tool has a vocabulary expanded to include many concepts from the domain of cultural heritage, there are many specialist concepts in the ASR transcription which are not included in its translation vocabulary. Approximately 3.5% of Dutch words were left untranslated (in addition to names). Some of these turned out to be important expressions, e.g. *rariteitenkabinet* 'cabinet of curiosities', which were in fact successfully retrieved by the systems for Run 1 and 2 (although ranked lower than desired).

The other main problem we encountered in Run 3 lay in the English Wikipedia and our limited experience concerning its structure. The downloadable dump includes a large number of pages that look like useful articles, but are in fact not. These articles include old articles set for deletion and meta-articles containing discussion of an existing, previous or future article. We were not aware of these articles during the initial development phase, but this had a significant impact on our results, about 18.5 % of the links returned in Run 3 proved to be invalid articles. Run 3′ reflects results where the English Wikipedia archive has been cleaned up to remove these irrelevant pages prior to indexing. As shown in Table 1, this cleanup produces a significant improvement in performance. A similar cleanup applied to the Dutch collection would produce a new ranking of Dutch documents. However, very few of the Dutch pages which would be deleted in cleanup are actually retrieved or have a link to English pages, and thus any changes in the Dutch archive will have no noticeable effect on evaluation of the overall system output.

5 Conclusions

In this paper we have outlined the two approaches used in our submissions to the Linking Task at VideoCLEF 2009. We found using the source language for retrieval to be more effective than switching to the target language in an early phase. This result may be different if translation of the query for the second method were to be improved. Both methods could be expected to benefit from the ongoing development of Wikipedia collections.

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, M., Newman, E. and Jones, G. J. F.: Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment. In: *Multilingual Information Access Evaluation Vol. II Multimedia Experiments.* Lecture Notes in Computer Science, Springer (2010)
2. The Lemur Toolkit. http://www.lemurproject.org/
3. Jones, G. J. F., Fantino, F., Newman, E., Zhang, Y.: 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, pages 34–41, (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/