A cocktail approach to the VideoCLEF'09 linking task

Stephan Raaijmakers Corné Versloot
Joost de Wit
TNO Information and Communication Technology
Delft, The Netherlands
{stephan.raaijmakers,corne.versloot, joost.dewit}@tno.nl

Abstract

In this paper, we describe the TNO approach to the Finding Related Resources or linking task of VideoCLEF09. Our system consists of a weighted combination of off-the-shelf and proprietary modules, including the Wikipedia Miner toolkit of the University of Waikato. Using this cocktail of largely off-the-shelf technology allows for setting a baseline for future approaches to this task.[1]

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*; I.3.1 [**Information Storage and retrieval**]: Content Analysis and indexing – *linguistic processing*

General Terms

Algorithms, Experimentation

Keywords

Wikipedia, Data Mining, Semantic Annotation, Word Sense Disambiguation

1 Introduction

The Finding Related Resources or linking task of VideoCLEF'09 consists of relating Dutch automatically transcribed TV speech to English Wikipedia content. For a total of 45 video episodes, a total of 165 anchors (speech transcripts) have to be linked to related Wikipedia articles. Technology emerging from this task will contribute to a better understanding of Dutch video for non-native speakers.

The TNO approach to this problem consists of a cocktail of off-the-shelf techniques. Central to our approach is the use of the Wikipedia Miner toolkit developed by researchers at the University of Waikato[2] (see [4]). The so-called *Wikifier* functionality of the toolkit detects Wikipedia topics from raw text, and generates cross-links from input text to a relevance-ranked list of Wikipedia pages.

We investigated two possible options for bridging the gap between Dutch input text and English Wikipedia pages: translating queries to English prior to the detection of Wikipedia topics, and translating Wikipedia topics detected in Dutch texts to English Wikipedia topics. In the latter case, the use of Wikipedia allows for an abstraction of raw queries to Wikipedia topics, for which

[1]This work is supported by the European IST Programme Project FP6-0033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

[2]See `http://wikipedia-miner.sourceforge.net`

the translation process in theory is less complicated and error prone. Specific to our approach is a weighted combination of various modules, and the use of a specially developed part-of-speech tagger for uncapitalized speech transcripts.

2 System setup

In this section, we describe the setup of our system. In subsections 2.1, 2.2 and 2.3, we describe the essential ingredients of our system. In subsection 2.4, we define a number of linking strategies based on these basic ingredients, which are combined into scenarios for our runs (section 3).

2.1 From Dutch to English

For the translation of Dutch text to English (and following [1]), we used the Yahoo! BabelFish translation service[3]. An example of the output of this service is given in Figure 1.

Figure 1: The result of Babelfish for a sample query.

Since people, organizations and locations often have entries in Wikipedia, accurate proper name detection is important for this task. Erroneous translation to English of Dutch names (e.g. 'Frans Hals' becoming 'French Neck') should be avoided. Proper name detection prior to translation allows for exempting the detected names from translation. A complicating factor is formed by the fact that the transcribed speech in the various broadcastings is in lowercase, which makes the recognition of proper names challenging, since important capitalization features can no longer be used. In order to address this problem, we trained a maximum entropy part-of-speech tagger: an instance of the Stanford tagger[4] (see [5]). The tagger was trained on a 700K part-of-speech tagged corpus of Dutch, after having decapitalized the training data. The feature space consists of a 5-cell bidirectional window addressing part-of-speech ambiguities and prefix and suffix features up to a size of 3.

[3] http://babelfish.yahoo.com/
[4] http://nlp.stanford.edu/software/tagger.shtml

2.2 A Dutch Wikifier

The imperfect English translation by Babel Fish was observed to be the main reason for erroneous Wikify results. In order to omit the translation step, we ported the English Wikifier of the Wikipedia Miner toolkit to Dutch, for which we used the Dutch Wikipedia dump and Perl scripts provided by developers of the Wikipedia Miner toolkit. The resulting Dutch Wikifier ('NL Wikifier' in Figure 3) has exactly the same functionality as the English version, but unfortunately contains a lot less pages than the English version (a factor 6 less). Even so, the translation process now is narrowed down to translating detected Wikipedia topics (the output of the Dutch Wikify step) to English Wikipedia topics. For the latter, we implemented a simple database facility (to which we shall refer with 'The English Topic Finder') that uses the cross-lingual links between topics in the Wikipedia database for carrying out the translation of Dutch topics to English topics.

An example of the output of the English and Dutch Wikifiers for the query in Figure 1 is given in Figure 2. The different rankings of the various detected topics are represented as a tag cloud

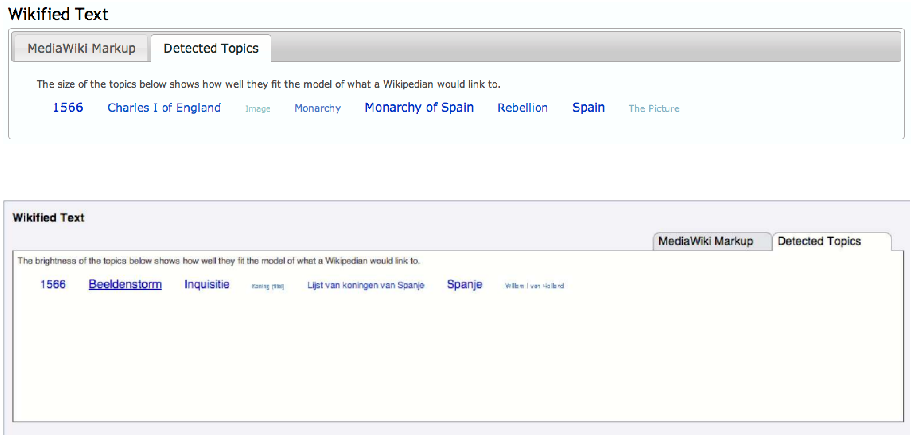

Figure 2: The result of the English and Dutch Wikifiers for a sample query.

with different font sizes, and can be extracted as numerical scores from the output.

2.3 Text retrieval

In order to be able to entirely by-pass the Wikipedia Miner toolkit, we deployed the Lucene search engine ([2]) for performing the matching of raw, translated text with Wikipedia pages. Lucene was used to index the Dutch Wikipedia with the standard Lucene indexing options. Dutch speech transcripts were simply provided to Lucene as a a disjunctive (OR) query, with Lucene returning the best matching Dutch Wikipedia pages for the query. The HTML of these pages was subsequently parsed in order to extract the English Wikipedia page references (which are indicated in Wikipedia, whenever present).

2.4 Linking strategies

The set of techniques just described leads to a total of four basic linking strategies.

Of the various combinatorial possibilities of these strategies, we selected 5 combinations for our submitted runs. The basic linking strategies are:

Strategy 1: proper names only (the top row in Figure 3) Following proper name recognition, a quasi-document is created that only consists of all recognized proper names. The Dutch Wikify tool is used to produce a ranked list of Dutch Wikipedia pages for this quasi-document.

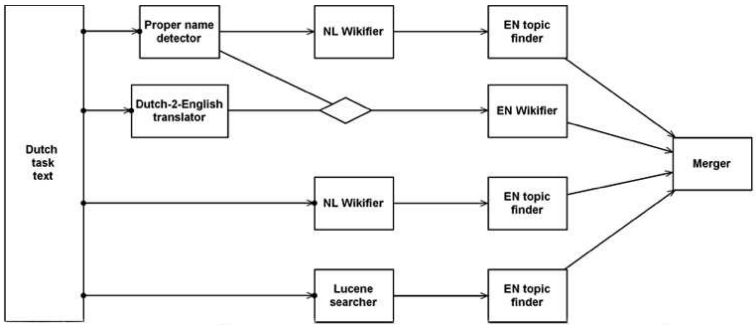

Figure 3: TNO system setup.

Subsequently, the topics of these pages are linked to English Wikipedia pages with the English Topic Finder.

Strategy 2: proper names preservation (second row in Figure 3) Dutch text is translated to English with Babelfish. Any proper names found in the part-of-speech tagged Dutch text are added to the translated text as untranslated text, after which the English Wikifier is applied, producing a ranked list of matching Wikipedia pages.

Strategy 3: topic to topic linking (3rd row from the top in Figure 3) The original Dutch text is wikified using the Dutch Wikify tool, producing a ranked list of Wikipedia pages. The topics of these pages are subsequently linked to English Wikipedia pages with the English Topic finder.

Strategy 4: text to page linking (bottom row in Figure 3) Lucene matches queries with Dutch Wikipedia pages. The English topic finder tries to find the corresponding English Wikipedia pages for the Dutch topics in the pages returned by Lucene. This strategy omits the use of the Wikifier and was used as a fall-back option, if none of the other modules delivered a result.

A thresholded merging algorithm removes any results below an estimated threshold and blends the remaining results into a single ordered list of Wikipedia topics, using estimated weights for the various sources of these results. Several different merging techniques were used for different runs; these will be discussed in subsection 3.

3 Run scenarios

In this section, we describe the configurations of the 5 runs we submitted. We were specifically interested in the effect of proper name recognition, the relative contributions of the Dutch and English Wikifiers, and the effect of full-text Babelfish translation as compared to a topic-to-topic translation approach.

Run 1

All four linking strategies were used to produce the first run. A weighted merger ('Merger' in Figure 3) was used to merge the results from the different strategies. The merger works as follows:

1. English Wikipedia pages referring to proper names are uniformly ranked before all other results.

2. The rankings produced by the second linking strategy ($rank_{EN}$) and third linking strategy ($rank_{DU}$) for any returned Wikipedia page p are combined according to the following scheme:

$$rank(p) = ((rank_{EN}(p) * 0.2) + (rank_{DU}(p) * 0.8)) * 1.4 \quad (1)$$

   The Dutch score was found to be more relevant than the English one (hence the 0.8 vs. 0.2 weights). The sum of the Dutch and English score is boosted with an additional factor of 1.4, awarding the fact that both linking strategies come up with the same result.

3. Pages found by linking strategy 2 but not by linking strategy 3 are added to the result and their ranking score is boosted with a factor of 1.1.

4. Pages found by linking strategy 3 but not by linking strategy 2 are added to the result (but their ranking score is not boosted).

5. If linking strategies 1 to 3 did not produce results, the results of linking strategy 4 are added to the result.

Run 2

Run 2 is the same as run 1 with the exception that linking strategy 1 is left out (no proper name recognition).

Run 3

Run 3 is similar to run 1, but does not boost results at the merging stage, and averages the rankings of the second and third linking strategy. This means that the weights used by the merger in run 1 (0.8, 0.2 and 1.4) are resp. 0.5, 0.5 and 1.0 for this run.

Run 4

Run 4 only uses linking strategy 1 and 3. This means that no translation from Dutch to English is performed. In the result set, the Wikipedia pages returned by linking strategy 1 are ordered before the results from linking strategy 2.

Run 5

Run 5 uses all linking strategies except linking strategy 1 (it omits proper name detection). In this run a different merging strategy is used:

1. If linking strategy 2 produces any results, add those to the final result set and then stop.

2. If linking strategy 2 produces no results, but linking strategy 3 does, add those to the final result and stop.

3. If none of the preceding linking strategies produces any results, add the results from linking strategy 3 to the final result set.

4 Results and discussion

For VideoCLEF'09, two groups submitted runs for the linking task: Dublin City University and TNO. Two evaluation methods were applied by the task organizers to the submitted results. A team of assessors first achieved consensus on a primary link (the most important or descriptive Wikipedia article), with a minimum consensus among 3 people. All queries in each submitted run

were scored for Mean Reciprocal Rank[5] for this primary link, as well as for recall. Subsequently, the annotators agreed on a set or related resources that necessarily included the primary link, in addition to secondary relevant links (minimum consensus of one person). Since this list of secondary links is non-exhaustive, for this measure only MRR is reported, and not recall.

Run	Recall	MRR
1	0.345	0.23
2	0.333	0.215
3	0.352	0.251
4	0.267	0.182
5	0.285	0.197
Average TNO	0.32	0.215

Table 1: Recall and MRR for the primary link evaluation by TNO. (Average DCU scores were 0.21 and 0.14, resp.)

Run	MRR
1	0.46
2	0.428
3	0.484
4	0.392
5	0.368
Average TNO	0.43

Table 2: MRR for the secondary link evaluation by TNO. (Average DCU score was 0.21.)

As it turns out, the unweighted combination of results (run 3) outperforms all other runs, followed by the thresholded combination (run 1). This indicates that the weights in the merging step are suboptimal. Omitting proper name recognition results in a noticeable drop of performance under both evaluation methods, underlining the importance of proper names for this task.

In addition to the recall and MRR scores, the assessment team distributed the graded relevance scores (see [3]) assigned to all queries. In Figure 4, we plotted the difference per query of the obtained averaged relevance score with the total average obtained relevance scores for both the Dublin City University (DCU) and TNO runs. For every video, we averaged the relevance scores of the hits reported by DCU and TNO. Subsequently, for every TNO run, we averaged relevance scores for every query, and measured the difference with the averaged DCU and TNO runs. It can be clearly seen that Run 1 and 3 obtain the best results, producing only a small amount of queries below the mean. Most of the relevance results obtained from these runs are around the mean, showing that from the perspective of relevance quality, our best runs produce average results.

5 Conclusions

In this contribution, we have taken a technological and off-the-shelf-oriented approach to the problem of linking Dutch transcripts to English Wikipedia pages. Using a blend of commonly available

[5]For a response $r = r_1, \ldots, r_Q$ to a ranking task, the MRR would be $MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i}$, with $rank_i$ the rank of answer r_i with respect to the correct answer.

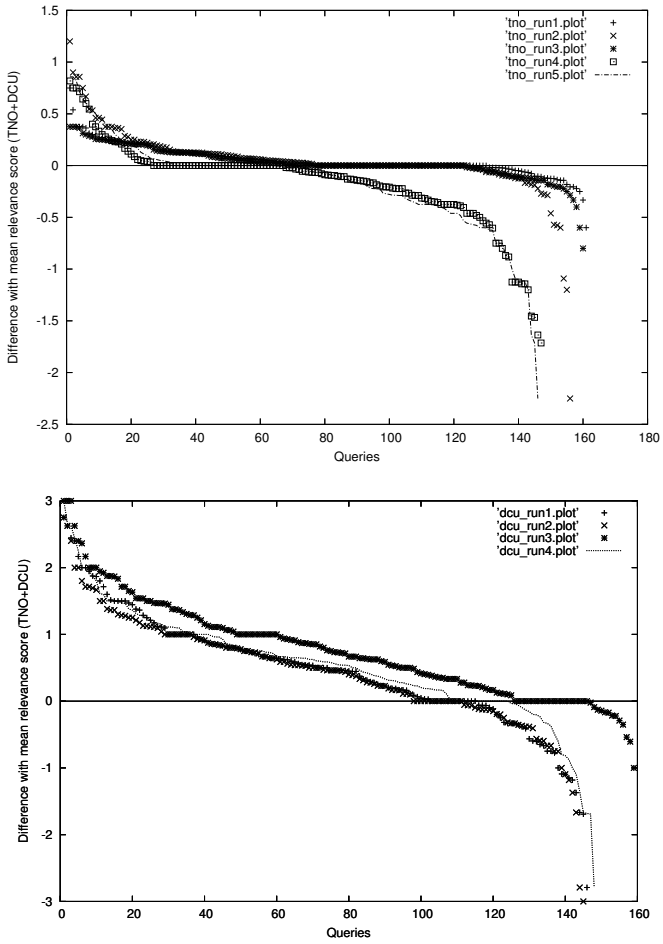

Figure 4: Difference plots of the various TNO (top figure) and DCU (bottom figure) runs compared to the averaged relevance scores of DCU and TNO.

software resources (Babelfish, the Waikato Wikipedia Miner Toolkit, Lucene, and the Stanford maximum entropy part-of-speech tagger), we demonstrated that an unweighted combination produces competitive results. We hope to have demonstrated that this low-entry approach can be used as a baseline level that can inspire future approaches to this problem. A more accurate estimation of weights for the contribution of several sources of information can be carried out in future benchmarks, now that the VideoClef annotators have produced ground truth ranking data.

References

[1] S. F. Adafre and Maarten de Rijke. Finding Similar Sentences across Multiple Languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69, 2006.

[2] Erik Hatcher and Otis Gospodnetic. *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA, 2004.

[3] Jaana Kekäläinen and Kalervo Järvelin. Using graded relevance assessments in IR evaluation. *J. Am. Soc. Inf. Sci. Technol.*, 53(13):1120–1129, 2002.

[4] David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge mining (CIKM'08)*, pages 509–518, New York, NY, USA, 2008. ACM.

[5] Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part–of–speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70, 2000.