

Topic Labeling of Multilingual Broadcast News in the Informedia Digital Video Library

Alexander G. Hauptmann, Danny Lee and Paul E. Kennedy

Abstract

The Informedia Digital Video Library Project includes a multilingual component for retrieval of video documents in multiple languages and a topic-labeling component for English video documents. We now extend this capability to English topic labeling of foreign-language broadcast-news stories. News stories are coarsely machine-translated into English, then assigned to a topic category using a K-nearest-neighbor algorithm. In preliminary tests on Croatian television news, topic assignment based on the best available machine translation technology showed performance only 8% worse (on a standard F-measure of performance) than that based on manual document translation. Using a phrase-based MT module the performance degradation was 31%.

1 The Informedia Digital Video Library

The Informedia Digital Library Project [1,2] allows full content indexing and retrieval of text, audio and video material, similar to what is available today for text only. To enable this access to video, speech recognition is used to provide a text transcript for the audio track, image processing determines scene boundaries, recognizes faces and allows for image similarity comparisons. Everything is indexed into a searchable digital video library [4,6], where users can submit queries and retrieve relevant news stories as results. *News-on-Demand* is a particular collection in the Informedia Digital Library that has served as a test-bed for automatic library creation techniques. As of July 1998, the Informedia project had about 1.3 terabytes of news video indexed and accessible online, with 1200 news broadcasts containing 24000 news stories.

The Informedia digital video library system has two distinct subsystems: the Library Creation System and the Library Exploration Client. The library creation system runs every night, automatically capturing, processing and adding current news shows to the library. It is during the library creation phase, that topics for news stories are automatically assigned to incoming stories. In [17], we described and evaluated tested a topic labeling component for the English language version of the Informedia Digital Video Library. During library exploration, the user can browse or search these stories and topics using the library exploration client. At 5 topics, the KNN-based system's recall was 0.49; and relevance was 0.48, with an F-measure at equal recall and precision of about 0.48.

2 Related Research on Topic Detection

The work reported here is similar in spirit to an approach reported by Schwartz [4], who classifies news stories into a static set using a Hidden Markov Model approach and found that to be somewhat better than a naïve Bayesian approach. Yang [7] also reports on other techniques, which try to cluster news stories into stories of similar topic content. This work differs in that the topic categories here are defined a priori, and do not change over with different data sets. We felt a fixed set of categories would better reflect the user needs than a clustering approach, which could yield different clusters on different days, depending on the contents of the corpus. We are extending the topic detection work and applying it in combination with machine translation techniques.

3 Multilingual Informedia

The Multilingual Informedia Project demonstrates a seamless extension of the Informedia approach to search and discovery across video documents in multiple languages. The new system performs speech recognition on

foreign language news broadcasts, segments it into stories and indexes the foreign data together with English news data from English language sources.

3.1 The Components of Multi-Lingual Informedia

There are three components in the Multilingual Informedia System [19] that differ significantly from the original Informedia system:

The speech recognizer recognizes a foreign language, specifically Croatian [9,10,11]. This component will not be described here.

In the first multi-lingual Informedia system, translingual broadcast retrieval was enabled by machine translation of an English query into the language(s) of the broadcasts, in our case Croatian. This enabled a search for equivalent words in a joint corpus of English and Croatian news broadcasts. A phrase-based translation module described in [19], provided the machine-translation capability. This translation module was also used to translate the complete broadcasts into English for some of the topic-detection experiments reported in this paper. In addition, for some of the current experiments, a version of the example-based machine translation system DIPLOMAT [18, 15] was used for “high-quality” translations of the news stories.

English topic labels for the foreign language news stories allow a user to identify a relevant story in the target language. In this paper, we will mostly describe this foreign language news topic classification component in detail.

3.2 The Informedia Translation Facility

The current version of the translation facility attempts to translate phrases it finds in a source-language text. The facility takes advantage of multi-word phrase entries in a machine-readable dictionary [16]. It uses a recursive procedure to search for dictionary entries corresponding to progressively smaller chunks of the input. The target-language equivalents of the chunks it finds get concatenated to form the output string. In general, this text-translation facility will work with any language pair so long as a bilingual machine-readable dictionary is available in the format the program understands.

The DIPLOMAT example-based machine translation system developed here at Carnegie Mellon University was also put to use for “high-quality” story translation from Croatian into English.

4 Foreign Language Topic Detection

After initial experiments with the Croatian news processed by the Multilingual Informedia system [19], it became clear that returning a foreign language result to the user was not sufficient. The users were unable to tell if a particular news clip was actually relevant to their query, or if it was returned due to poor query translation or inadequate information retrieval techniques. To allow the user at least some judgment about the returned stories, we attempted to label each Croatian news story with an English-language topic.

The topic identification was done using the Informedia translation facility to translate the whole story into English words. This translation became the *topic query*. Separately, we had indexed about 35000 English language news stories, which had manually assigned topics assigned to them. Using the SMART information retrieval system, we now used the translated *topic query* to retrieve the most relevant 10 labeled English stories. Each of the topics for the labeled stories that were retrieved was weighted by its relevance to the *topic query* and the weights for each topic were summed. The most favored topics, above a threshold, were then used to provide a topic label for the Croatian news story. This topic label allows the user to identify the general topic area of an otherwise incomprehensible foreign language text and determine if it is relevant at least in the topic area.

5 Experiments

For initial (non-translingual) training and testing of the topic detection system English data was taken from a set of CD-ROMs of broadcast news transcripts, published by Primary Source Media [8]. The online Informedia system uses actual broadcast video, for which no manual topic labels are available; however, the data is of the same type as on the CD-ROM.

From this CDROM, we used 34671 news stories from 1995 as training data. Each of the news stories had one or more topic labels associated with it. Of these topic labels, we selected the top 3178 unique topics, which occurred at least 10 times in the whole corpus. Topics with fewer instances were viewed as idiosyncratic and ignored in the experiments. To test the accuracy of the topic assignment, 24 news stories broadcast in 1996 by Croatian Television in Zagreb were manually transcribed. A typical story, which was manually translated into English, is given in the following paragraph:

“Palestinian leader Yassir Arafat gave an announcement in Cairo, after his talk with the Egyptian President Mubarak and the main secret agent Magid of the Arab League, that inspite of the Israelites protests he intends to proclaim independent Palestinian state in the area of the West Coast and Gaze. Palestinian State is not Israel's problem, but the Arabian and international, says Arafat who believes all contracts up to date are only temporary, reports Reuter. The proclamation for the state should be the crowning of the peace talks. It is not necessary to point to the Israelites opposition regarding this matter”

For the above story, human transcribers marked the following topic labels as relevant: *“Administration; Arafat, Yasir; Israel; Jewish-Arab relations; Middle East; Middle East peace negotiations; Occupied territories; Palestine Liberation Organization; Palestinian Arabs; Palestinian self-rule areas”*.

The same story translated using DIPLOMAT and after stopword removal and stemming was:

“palestinian leader yasser arafat declar talk league proclaim independ countri territori zapadne palestinian stat question arabic report theodor agreements probat countri not peac negotiations opposit point view expire”

The phrase-based translation system produced the following translation (again after stopword removal and stemming):

“palestinian commander's ship head leader beat fight flow pour strike self-adjusting consisting spite announce declare proclaim consisting country government state question consisting beat fight flow pour strike back intended made. probate proclamation beat fight flow pour strike opposition elapse emphasize expire flow hang hoist leak placard post raise run underline”

5.1 Method

The algorithm for the topic-labeling module was based on a k-nearest neighbor (KNN) strategy [6,7]. The process is split into a training and a classification phase. The training phase only occurs once, but each incoming story document must be classified separately.

During the **training phase**, the system received as input a set of 34671 broadcast news stories from the year 1995, which already had (manually) assigned topics. On average, each news story document had 5.48 multiple topics assigned to it. Each news story was preprocessed which removed stop words, and each word was converted into its stemmed root form. The entire set of documents was then indexed using a vector space search engine (SMART) [3]. The weighting scheme used in SMART was “mnc”.

During the **classification phase** each new, unclassified news story was also preprocessed to remove stop words and convert words into their root stems. The unclassified document was then vectorized into the SMART vector space using the “lrc” weighting scheme. A distance between the unclassified news story vector and each of the training story vectors was computed using the cosine similarity measure.

The 10 top ranked training documents were selected based on their close similarity to the unclassified news story. Every topic assigned to these top-10 training documents is assigned the same similarity score as the training document itself. The similarity scores of multiple instances of the same topic in several top-10 stories were summed for the topic. The final topic similarity score was used as the topic relevance score, providing a relevance of the topic to the new, previously unclassified document. The top relevant topics above a threshold relevance of 0.8 were selected as the topics to label the new story. These topic labels were then added to the indexed Informedia News-on-Demand database, which then allows searching on topics, as well as browsing.

5.2 Results

Recall and relevance were measured for an independent test set of 24 news stories. Each story had an average of 9.6 relevant topics. We compared those topic labels with the topics generated by the KNN method. In other words, of the topics that the KNN method generated, how many were the same as the ones assigned by a human (precision) and how many of the human assigned topics did we correctly assign using the KNN method (recall).

To measure our topic classification effectiveness, we use a widely accepted metric called the F-measure, which is defined as:

$$F(n) = \frac{(n^2 + 1) * Precision * Recall}{(n^2 * Precision) * Recall}$$

where:

Recall (Sensitivity) = Number of relevant items retrieved / Number of relevant items in database

Precision (Positive Predictive Value) = Number of relevant items retrieved / Number of items retrieved.

If $n=1$, then precision and recall are weighted as equally important, which is our assumption for the current evaluation.

In the independent test set from Croatian television news, an evaluation which combines equal precision and recall yielded an $F(1)$ measure value of 0.46 on manually translated Croatian television news stories. The machine translation experiment showed only a slight (8%) decrease in topic classification accuracy from 0.46 to 0.43, using our best available machine translation technology. Using phrase-based translation, the decrease was 31% to an F-measure value of 0.32. In all cases, even the manual translation, the system performed worse than the English topic assignment on English language news stories reported in [17], which had an F-measure value at equal precision and recall of over 0.48.

5.3 Conclusions

In summary, we found the approach to be promising and these initial results to be encouraging. In particular, a high-quality machine translation provided topic assignment results comparable to perfect translations. The quick-and-dirty phrase translation system, however, showed noticeable degradation in the topic assignment.

As a near-term next step we plan to continue this evaluation of the KNN topic classification approach in the multilingual news setting. Specifically, we intend to take speech recognizer generated transcripts, which are then automatically translated and compare the classification performance under the manual, high-quality (DIPLOMAT example-based machine translation) and low-quality (phrase-based) translation schemes.

In general, there are drawbacks to the use of manual-generated news topics from a limited epoch, which directly reflect current issues of the time period (E.g. the Princess of Wales or O.J. Simpson figured prominently as topic categories at particular times). In the long term, we would like to shift away from the ad-hoc set of topics used in the broadcast news transcript CDROM, to a carefully defined set of hierarchical categories. Possible candidates are the Dewey Decimal Classification system or the Library of Congress Classification Scheme, which are popular in libraries around the world. We also would like to provide a tight

integration of the topic classification into the browsing and navigation component of the Multilingual Informedia system, instead of merely allowing users to browse or search for these topics.

6 Acknowledgments

This paper is based on work supported by the National Science Foundation, DARPA and NASA under NSF Cooperative agreement No. IRI-9411299. Thanks to Peter Scheytt for his help with the Croatian news data.

7 References

1. Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H.", "Informedia Digital Video Library", *Communications of the ACM*, 38 (4), April 1994, pp. 57-58.
2. The Informedia Digital Video Library Project <http://www.informedia.cs.cmu.edu/>
3. Salton, G., Ed, "The SMART Retrieval System", Prentice-Hall, Englewood Cliffs, 1971.
4. Schwartz, R., Imai, T., Kubala, F., Nguyen, L., and Makhoul, J., A Maximum Likelihood Model for Topic Classification in Broadcast News, Eurospeech-97 – 5th European Conference on Speech Communication and Technology, Rhodes, Greece, September 1997.
5. Thompson, R., Shafer, K., and Vizine-Goetz, D., Evaluating Dewey Concepts as a Knowledge Base for Automatic Subject Assignment, http://orc.rsch.oclc.org:6109/eval_dc.html
6. Yang, Y., Carbonell, J. G., Allan, J., Yamron, J. Topic Detection and Tracking: Detection-Task, Project report on the TDT Workshop, Oct 1997.
7. Yang, Y., An Evaluation of statistical approach to text categorization. Technical Report CMU-CS-97-127, Computer Science Department, Carnegie Mellon University, 1997.
8. Primary Source Media, Broadcast News CDROM, Woodbridge, CT, 1995, 1996.
9. Geutner, P., Finke, M., Scheytt, P., Waibel, A., and Wactlar, H., "Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexicon Adaptation." In *BNTUW-98 Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne VA, February 1998.
10. Geutner, P., Finke, M., Scheytt, P., Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News, Technical Report, Carnegie Mellon University, Pittsburgh, PA, CMU-LTI-97-155, December 1997.
11. Scheytt, P., Finke, M., Geutner, P., Speech Recognition on Serbo-Croatian Dictation and Broadcast News Data, Technical Report, Carnegie Mellon University, Pittsburgh, PA, CMU-LTI-97-154, December 1997.
12. Witbrock, M.J., and Hauptmann, A.G., "Speech Recognition and Information Retrieval", Proceedings of the 1997 DARPA Speech Recognition Workshop, Chantilly, VA, February 2-5, 1997.
13. Wactlar, H.D., Kanade, T., Smith, M.A. and Stevens, S.M. "Intelligent Access to Digital Video: Informedia Project". *IEEE Computer*, 29(5) May 1996, p.p. 46-52.
14. Hauptmann, A.G. and Witbrock, M.J., *Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval*, In Maybury, M. (ed.), "Intelligent Multimedia Information Retrieval", AAAI Press, 1997.

15. Carbonell, J., Yang, Y., Frederking, R., Brown, R.D., Geng, Y., and Lee, D. "Translingual Information Retrieval: A Comparative Evaluation". In Proceedings of Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97).
16. Brown, R.D., "Automated Dictionary Extraction for ``Knowledge-Free" Example-Based Translation". In Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation, Santa Fe, July 23-25, 1997.
17. Hauptmann, A.G., and Lee, D., Topic Labeling of Broadcast News Stories in the Infromedia Digital Video Library Digital Libraries '98 - The Third ACM Conference on Digital Libraries, Pittsburgh, PA, June, 1998.
18. Frederking, R., Rudnicky, A., and Hogan, C. Interactive Speech Translation in the DIPLOMAT Project. Spoken Language Translation workshop of the Association for Computational Linguistics, ACL-97. Madrid, Spain. 1997.
19. Hauptmann, A.G., Scheytt, P., Wactlar, H.D., and Kennedy, P.E., Multi-Lingual Infromedia: A Demonstration of Speech Recognition and Information Retrieval across Multiple Languages, BNTUW-98 Proceedings of the DARPA Workshop on Broadcast News Understanding Systems, Lansdowne, VA, February 1998.