Automatic Title Generation using EM

Paul E. Kennedy
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420
Email: pkennedy@cs.cmu.edu

Alexander G. Hauptmann
Dept. of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
http://www.informedia.cs.cmu.edu/
Email: alex@cs.cmu.edu

ABSTRACT
Our prototype automatic title generation system inspired by statistical machine-translation approaches [1] treats the document title like a translation of the document. Titles can be generated without extracting words from the document. A large corpus of documents with human-assigned titles is required for training title "translation" models. On an f1 evaluation score our approach outperformed another approach based on Bayesian probability estimates [7].

KEYWORDS: document summarization, title assignment

THE APPROACH
Extractive summarization is the most common approach to generate titles or short summaries of text data [3, 5]. The interesting phrases are usually determined through a variant of a TFIDF (Term Frequency by Inverse Document Frequency) word score for each document sentence. Highly interesting phrases are included in the headline summary. Our approach is non-extractive; a summary does not have to consist of phrase snippets taken from the document. For a statistical approach to summarization using naïve Bayesian estimates instead of an Estimation/Maximization algorithm (EM), see Witbrock and Mittal [7].

The IBM machine translation approach, which inspired our title summarization work, uses a source-channel model: Given a French source language string **f**, find the English target language text string **e** most likely to represent the translation that produced **f**, i.e., find

$$\text{Argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) = \text{Argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})p(\mathbf{e}) \quad \text{[Bayes]}$$

By analogy, our title generation system generates a title for a document by estimating

$$\text{Argmax}_{\text{title}}\, p(\text{title}|\text{document}) = \text{Argmax}_{\text{title}}\, (\text{document}|\text{title})p(\text{title})$$

[2] used an English language model to estimate the prior probabilities p(**e**). Similarly, to estimate p(title), we use a standard trigram language model to define a space of possible titles and their prior probabilities.

To estimate p(**f**|**e**) the IBM researchers developed statistical models of alignments, i.e. the various ways words or phrases in an English sentence might translate into corresponding words or phrases in a French sentence. We have emulated the simplest of IBM's models, (Model 1 [2]), in order to estimate p(document|title). This model treats the title and document as a "bag of words". For a given pair of words, one from the title vocabulary and one from the document vocabulary, this model simply estimates the probability that the document word appears in a document given that the title word appears in the corresponding title. Thus the model consists of a list of document-word/title-word pairs, with a probability assigned to each. For a pair to be in the list there must have been an actual document/title pair in the training corpus where the title word occurs in the title and the document word occurs in the document. The probabilities are estimated in multiple iterations using the EM algorithm. For details of this approach we refer the reader to the well-known paper from IBM [2], but we outline the essential steps here. The discussion uses the following key:

e title word (English word) f doc word (French word)
e title (English sentence) **f** document (French sentence)

1. Each word in the title maps to one or more words in the document for a given alignment. Each document word maps to 0 or 1 title words. If the former, the document word maps to a "null" title word.

2. All possible combinations of word correspondence between document and title are allowed and equally probable. The title and document lengths are independent; for document length m, $p(m|\mathbf{e}) = \varepsilon$ some small constant for all m, **e**.

3. For a given title and document, if l is the length of the title, there are $(l + 1)^m$ possible alignments. The probability of each alignment is then $(l + 1)^{-m}$.

4. The model estimates a fixed "translation probability" t(f|e) for each French/English (document/title) word pair.

5. Given a document and a title **f** and **e**, it can be shown that

$$P(\mathbf{f}|\mathbf{e}) = \varepsilon\,(l+1)^{-m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i) \qquad (1)$$

where **f** is the string of words $f_1 f_2 \ldots f_m$ and **e** is the string of words $e_0 e_1 \ldots e_l$, with e_0 the "null" word.

The t(f|e)'s are estimated using the EM algorithm. At each iteration, the re-estimation uses the formulas

$$t_{new}(f|e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)}) \quad (2)$$

$$c(f|e; \mathbf{f}, \mathbf{e}) = t(f|e)[\sum_{i=0}^{l} t(f|e_i)]^{-1} [cnt(\mathbf{e}, e)][cnt(\mathbf{f}, f)] \quad (3)$$

where $\mathbf{f}^{(1)}$, $\mathbf{e}^{(1)}$, $\mathbf{f}^{(2)}$, $\mathbf{e}(2)$, …, $\mathbf{f}^{(S)}$, $\mathbf{e}^{(S)}$ are the document/title pairs in the training corpus, cnt(**e**, e) is the number of times e appears in **e**, cnt(**f**, f) is the number of times f appears in **f**, and λ_e is the normalization factor required to make t_{new} a probability distribution. Note that

$$\lambda_e = \sum_{s=1}^{S} \sum_f c(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)}) \quad (4)$$

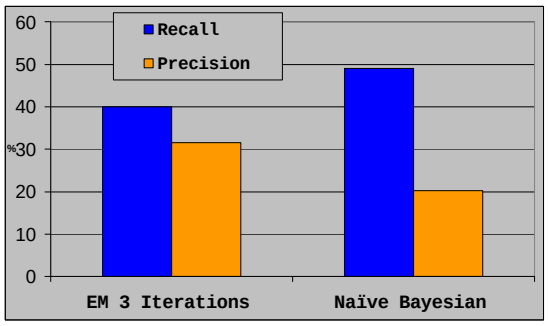
The EM algorithm converges to a global maximum in this model. Thus the initial values for the t(f|e)'s can be set arbitrarily (as long as they are normalized and not 0).

EXPERIMENTS

To evaluate our title generation approach, we trained a word-pair model P(dw|tw) for 3 iterations using the approach outlined above on a corpus of 40000 transcripts of broadcast-news stories with human-assigned titles and also built a standard trigram language model, as P(title), from just the titles in the corpus.

Using a held-out test set of 100 news stories, we selected the top 50 title words from each document that maximized $\sum_{dw \in doc} P(dw|tw)$ where dw denotes a document word and tw likewise denotes a title word. Recall and precision were computed as the percentage of words in the original (manual) reference title compared to the automatically generated list of the top 50 candidate title words. f1 was computed from these as f1 = 2pr/(p + r).

EM at 3 iterations (precision 40%, recall 31.5%, f1 .352) compares favorably with a Bayesian approach [7] (precision 20.2%, recall 49%, f1 .286) as shown below.



To create a linearized "English-like" title, a lattice was formed consisting of a regular set of 6 columns, each column being a copy of the top 50 list of title word candidates with corresponding probabilities. The lattice-rescorer from [6] is run with this lattice and the trigram language model for titles as input. The output of the lattice rescorer is taken as the finished title subject to a procedure to eliminate repeated words therefrom as in [7].

In the following sample results, the "**Ref**" title was human-generated in the corpus. "**Extractive**" is the title generated by Informedia [4] using TFIDF phrase extraction. The **Bayes** titles were generated using our implementation of [7]. Finally, titles generated through our **EM** approach are listed for comparison at the bottom.

Ref: MARKET OUTLOOK FOR THE NEW YEAR
Extractive: HARD ASSET PLAY WELL I'VE OWNED T S FAIRLY HEAVILY FOR COURSE OF ABOUT GOLD YOU WOULDN'T GO FOR T S YOU KNOW SORT OF …
Bayes: FINANCIAL NEWS OF THE STOCK MARKET
EM: STOCK MARKET STRATEGIST DISCUSSES WALL STREET

Ref: TURBULENCE INVESTIGATION
Extractive: O K LISA SIGN O K LISA THANKS VERY LISA WE'RE ALSO COMING UP THIS HALF HOUR IN OUR REPORT DRINKING CAN YOU JUDGE DRUNKENNESS
Bayes: NEW YORK NEWS OF THE DAY
EM: TWA FLIGHT 800 CRASH

Ref: IBM ANNOUNCES BREAKTHROUGH
Extractive: ANNOUNCEMENT FROM SHOWS THAT DISC DRIVES THAT STORE COMPUTER DATA ARE AL SO ON FAST TRACK AND TECHNOLOGY THAT SHOULD BECOME AV...
Bayes: NEW TECHNOLOGY NEWS OF THE HOUSE
EM: EFFORTS SAVE RESIDENTS PREPARE SET UNDERWAY

Based on these 3 examples, no strong qualitative statements can be made. While the approach feels promising and is theoretically appealing, further work is clearly needed.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under Cooperative Agreement IIS-9817496. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The approach was suggested by Professor John Lafferty of the Carnegie Mellon University School of Computer Science.

REFERENCES

1. Brown, Cocke, Della-Pietra, Della-Pietra, Jelinek, Lafferty, Mercer, Roossin, A Statistical Approach to Machine Translation, *Computational Linguistics*, 16(2) June 1990.
2. Brown, Della Pietra, Della Pietra, Mercer, The Mathematics of Machine Translation: Parameter Estimation, *Computational Linguistics* 19(2), June 1993.
3. Hovy, E. and Lin, C.Y., Automated text summarization in SUMMARIST. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 18–24, Madrid, Spain, 1997.
4. Informedia Digital Video Library Project, CMU 1999, http://www.informedia.cs.cmu.edu
5. Salton, G., Singhal, A., Mitra, M., and Buckley, C., Automatic text structuring and summary. Info. Proc. and Management, 33(2):193–207, March 1997.
6. K. Seymore, S. Chen, S.J. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer. The 1997 CMU Sphinx-3 English Broadcast News transcription system. Proceedings of the DARPA Speech Recognition Workshop, 1998. http://www.speech.cs.cmu.edu
7. Witbrock, M.J. and Mittal, V.O, Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries, in Proc. SIGIR 99 Research and Development in Information Retrieval (Berkeley, August 15-19, 1999), ACM Press, pp. 315-316.