An Unsupervised Approach to Biography Production using Wikipedia

Fadi Biadsy,[†] Julia Hirschberg[†] and Elena Filatova[*]

[†]Department of Computer Science
Columbia University, New York, NY 10027, USA
{fadi,julia}@cs.columbia.edu

[*]InforSense LLC
Cambridge, MA 02141, USA
efilatova@inforsense.com

Abstract

We describe an unsupervised approach to multi-document sentence-extraction based summarization for the task of producing biographies. We utilize Wikipedia to automatically construct a corpus of biographical sentences and TDT4 to construct a corpus of non-biographical sentences. We build a biographical-sentence classifier from these corpora and an SVM regression model for sentence ordering from the Wikipedia corpus. We evaluate our work on the DUC2004 evaluation data and with human judges. Overall, our system significantly outperforms all systems that participated in DUC2004, according to the ROUGE-L metric, and is preferred by human subjects.

1 Introduction

Producing biographies by hand is a labor-intensive task, generally done only for famous individuals. The process is particularly difficult when persons of interest are **not** well known and when information must be gathered from a wide variety of sources. We present an automatic, unsupervised, multi-document summarization (MDS) approach based on extractive techniques to producing biographies, answering the question “Who is X?”

There is growing interest in automatic MDS in general due in part to the explosion of multilingual and multimedia data available online. The goal of MDS is to automatically produce a concise, well-organized, and fluent summary of a set of documents on the same topic. MDS strategies have been employed to produce both generic summaries and query-focused summaries. Due to the complexity of text generation, most summarization systems employ sentence-extraction techniques, in which the most relevant sentences from one or more documents are selected to represent the summary. This approach is guaranteed to produce grammatical sentences, although they must subsequently be ordered appropriately to produce a coherent summary.

In this paper we describe a sentence-extraction based MDS procedure to produce biographies from online resources automatically. We make use of Wikipedia, the largest free multilingual encyclopedia on the internet, to build a biographical-sentence classifier and a component for ordering sentences in the output summary. Section 2 presents an overview of our system. In Section 3 we describe our corpus and in Section 4 we discuss the components of our system in more detail. In Section 5, we present an evaluation of our work on the Document Understanding Conference of 2004 (DUC2004), the biography task (task 5) test set. In Section 6 we compare our research with previous work on biography generation. We conclude in Section 7 and identify directions for future research.

2 System Overview

In this section, we present an overview of our biography extraction system. We assume as input a set of documents retrieved by an information retrieval engine from a query consisting of the name of the person for whom the biography is desired. We further assume that these documents have been tagged with Named Entities (NE)s with coreferences resolved

Proceedings of ACL-08: HLT, pages 807–815,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

using a system such as NYU's 2005 ACE system (Grishman et al., 2005), which we used for our experiments. Our task is to produce a concise biography from these documents.

First, we need to select the most 'important' biographical sentences for the target person. To do so, we first extract from the input documents all sentences that contain some reference to the target person according to the coreference assignment algorithm; this reference may be the target's name or a coreferential full NP or pronominal referring expression, such as *the President* or *he*. We call these sentences *hypothesis sentences*. We hypothesize that most 'biographical' sentences will contain a reference to the target. However, some of these sentences may be irrelevant to a biography; therefore, we filter them using a binary classifier that retains only 'biographical' sentences. These biographical sentences may also include redundant information; therefore, we cluster them and choose one sentence from each cluster to represent the information in that cluster. Since some of these sentences have more salient biographical information than others and since manually produced biographies tend to include information in a certain order, we reorder our summary sentences using an SVM regression model trained on biographies. Finally, the first reference to the target person in the initial sentence in the reordering is rewritten using the longest coreference in our hypothesis sentences which contains the target's full name. We then trim the output to a threshold to produce a biography of a certain length for evaluation against the DUC2004 systems.

3 Training Data

One of the difficulties inherent in automatic biography generation is the lack of training data. One might collect training data by manually annotating a suitable corpus containing biographical and non-biographical data about a person, as in (Zhou et al., 2004). However, such annotation is labor intensive. To avoid this problem, we adopt an unsupervised approach. We use Wikipedia biographies as our corpus of 'biographical' sentences. We collect our 'non-biographical' sentences from the English newswire documents in the TDT4 corpus.[1] While each corpus may contain positive and negative examples, we assume that most sentences in Wikipedia biographies are biographical and that the majority of TDT4 sentences are non-biographical.

3.1 Constructing the Biographical Corpus

To automatically collect our biographical sentences, we first download the xml version of Wikipedia and extract only the documents whose authors used the Wikipedia biography template when creating their biography. There are 16,906 biographies in Wikipedia that used this template. We next apply simple text processing techniques to clean the text. We select at most the first 150 sentences from each page, to avoid sentences that are not critically important to the biography. For each of these sentences we perform the following steps:

1. We identify the biography's subject from its title, terming this name the 'target person.'

2. We run NYU's 2005 ACE system (Grishman et al., 2005) to tag NEs and do coreference resolution. There are 43 unique NE tags in our corpora, including *PER_Individual*, *ORG_Educational*, and so on, and TIMEX tags for all dates.

3. For each sentence, we replace each NE by its tag name and type ([name-type_subtype]) as assigned by the NYU tagger. This modified sentence we term a *class-based/lexical sentence*.

4. Each non-pronominal referring expression (e.g., *George W. Bush*, *the US president*) that is tagged as coreferential with the target person is replaced by our own [TARGET_PER] tag and every pronoun *P* that refers to the target person is replaced by [TARGET_*P*], where *P* is the pronoun itself. This allows us to generalize our sentences while retaining a) the essential distinction between this NE (and its role in the sentence) and all other NEs in the sentence, and b) the form of referring expressions.

5. Sentences containing no reference to the target person are assumed to be irrelevant and removed from the corpus, as are sentences with

[1] http://projects.ldc.upenn.edu/TDT4

fewer than 4 tokens; short sentences are unlikely to contain useful information beyond the target reference.

For example, given sentences from the Wikipedia biography of Martin Luther King, Jr. we produce class-based/lexical sentences as follows:

Martin Luther King, Jr., was born on January 15, 1929, in Atlanta, Georgia. He was the son of Reverend Martin Luther King, Sr. and Alberta Williams King. He had an older sister, Willie Christine (September 11, 1927) and a younger brother, Albert Daniel.

[TARGET_PER], was born on [TIMEX], in [GPE_PopulationCenter]. [TARGET_HE] was the son of [PER_Individual] and [PER_Individual]. [TARGET_HE] had an older sister, [PER_Individual] ([TIMEX]) and a younger brother, [PER_Individual].

3.2 Constructing the Non-Biographical Corpus

We use the TDT4 corpus to identify non-biographical sentences. Again, we run NYU's 2005 ACE system to tag NEs and do coreference resolution on each news story in TDT4. Since we have no target name for these stories, we select an NE tagged as *PER_Individual* at random from all NEs in the story to represent the target person. We exclude any sentence with no reference to this target person and produce class-based/lexical sentences as above.

4 Our Biography Extraction System

4.1 Classifying Biographical Sentences

Using the biographical and non-biographical corpora described in Section 3, we train a binary classifier to determine whether a new sentence should be included in a biography or not. For our experiments we extracted 30,002 sentences from Wikipedia biographies and held out 2,108 sentences for testing. Similarly. we extracted 23,424 sentences from TDT4, and held out 2,108 sentences for testing. For each sentence, we then extract the frequency of three class-based/lexical features — unigram, biagram, and trigram — and two POS features — the frequency of unigram and bigram POS. To reduce the dimensionality of our feature space, we first sort the features in decreasing order of Chi-square statistics computed from the contingency tables of the observed frequencies from the training data. We then take the highest 30-80% features, where the number of features used is determined empirically for each feature type. This process identifies features that significantly contribute to the classification task. We extract 3K class-based/lexical unigrams, 5.5K bigrams, 3K trigrams, 20 POS unigrams, and 166 POS bigrams.

Classifier	*Accuracy*	*F-Measure*
SVM	87.6%	0.87
M. naïve Bayes	84.1%	0.84
C4.5	81.8%	0.82

Table 1: Binary classification results: Wikipedia biography class-based/lexical sentences vs. TDT4 class-based/lexical sentences

Using the training data described above, we experimented with three different classification algorithms using the Weka machine learning toolkit (Witten et al., 1999): multinomial naïve Bayes, SVM with linear kernel, and C4.5. Weka also provides a classification confidence score that represents how confident the classifier is on each classified sample, which we will make use of as well.

Table 1 presents the classification results on our 4,216 held-out test-set sentences. These results are quite promising. However, we should note that they may not necessarily represent the successful classification of biographical vs. non-biographical sentences but rather the classification of Wikipedia sentences vs. TDT4 sentences. We will validate these results for our full systems in Section 5.

4.2 Removing Redundant Sentences

Typically, redundancy removal is a standard component in MDS systems. In sentence-extraction based summarizers, redundant sentences are defined as those which include the same information without introducing new information and identified by some form of lexically-based clustering. We use an implementation of a single-link nearest neighbor clustering technique based on stem-overlap (Blair-Goldensohn et al., 2004b) to cluster the sentences classified as biographical by our classifier, and then select the sentence from each cluster that maximizes the confidence score returned by the classifier as the representative for that cluster.

4.3 Sentence Reordering

It is essential for MDS systems in the extraction framework to choose the order in which sentences should be presented in the final summary. Presenting more important information earlier in a summary is a general strategy for most domains, although importance may be difficult to determine reliably. Similar to (Barzilay and Lee, 2004), we automatically learn how to order our biographical sentences by observing the typical order of presentation of information in a particular domain. We observe that our Wikipedia biographies tend to follow a general presentation template, in which birth information is mentioned before death information, information about current professional position and affiliations usually appear early in the biography, and nuclear family members are typically mentioned before more distant relations. Learning how to order information from these biographies however would require that we learn to identify particular types of biographical information in sentences.

We directly use the position of each sentence in each Wikipedia biography as a way of determining where sentences containing similar information about different target individuals should appear in their biographies. We represent the absolute position of each sentence in its biography as an integer and train an SVM regression model with RBF kernel, from the class/lexical features of the sentence to its position. We represent each sentence by a feature vector whose elements correspond to the frequency of unigrams and bigrams of class-based items (e.g., GPE, PER) (cf. Section 3) and lexical items; for example, the unigrams *born*, *became*, and [*GPE_State-or-Province*], and the bigrams *was born*, [*TARGET_PER*] *died* and [*TARGET_PER*] *joined* would be good candidates for such features.

To minimize the dimensionality of our regression space, we constrained our feature choice to those features that are important to distinguish biographical sentences, which we term **biographical terms**. Since we want these biographical terms to impact the regression function, we define these to be phrases that consist of at least one lexical item that occurs in many biographies but rarely more than once in any given biography. We compute the biographical term score as in the following equation:

$$bio\_score(t) = \frac{\mid D_t \mid}{\mid D \mid} \cdot \frac{\sum_{d \in D_t} (1 - \frac{n(t)_d}{\max_t (n(t)_d)})}{\mid D \mid} \quad (1)$$

where D is the set of 16,906 Wikipedia biographies, $n(t)_d$ is the number of occurrences of term t in document d, and $D_t = \{d \in D : t \in d\}$. The left factor represents the document frequency of term t, and the right factor calculates how infrequent the term is in each biography that contains t at least once.[2] We order the unigrams and bigrams in the biographies by their biographical term scores and select the highest 1K unigrams and 500 bigrams; these thresholds were determined empirically.

4.4 Reference Rewriting

We observe that news articles typically mention biographical information that occurs early in Wikipedia biographies when they mention individuals for the first time in a story (e.g. *Stephen Hawking, the Cambridge University physicist*). We take advantage of the fact that the coreference resolution system we use tags full noun phrases including appositives as part of NEs. Therefore, we initially search for the sentence that contains the longest identified NE (of type PER) that includes the target person's full name and is coreferential with the target according to the reference resolution system; we denote this NE *NE-NP*. If this sentence has already been classified as a biographical sentence by our classifier, we simply boost its rank in the summary to first. Otherwise, when we order our sentences, we replace the reference to the target person in the first sentence by NE-NP. For example, if the first sentence in the biography we have produced for Jimmy Carter is *He was born in 1947* and a sentence not chosen for inclusion in our biography *Jimmy Carter, former U.S. President, visited the University of California last year.* contains the NE-NP, and *Jimmy Carter* and *He* are coreferential, then the first sentence in our biography will be rewritten as *Jimmy Carter, former U.S. President, was born in 1947.* Note that, in the evaluations presented in Section 5, sentence order was modified by this process in only eight summaries.

5 Evaluation

To evaluate our biography generation system, we use the document sets created for the biography evalua-

[2] We considered various approaches to feature selection here, such as comparing term frequency between our biographical and non-biographical corpora. However, terms such as *killed* and *died*, which are useful biographical terms, also occur frequently in our non-biographical corpus.

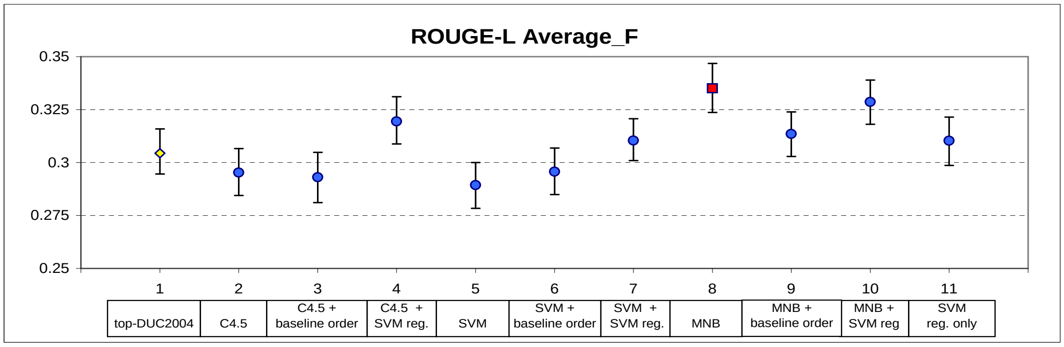

Figure 1: Comparing our approaches against the top performing system in DUC2004 according to ROUGE-L (diamond).

tion (task 5) of DUC2004.[3] The task for systems participating in this evalution was " *Given each document cluster and a question of the form "Who is* X*?", where* X *is the name of a person or group of people, create a short summary (no longer than 665 bytes) of the cluster that responds to the question.*" NIST assessors chose 50 clusters of TREC documents such that all the documents in a given cluster provide at least part of the answer to this question. Each cluster contained on average 10 documents. NIST had 4 human summaries written for each cluster. A baseline summary was also created for each cluster by extracting the first 665 bytes of the most recent document in the cluster. 22 systems participated in the competition, producing a total of 22 automatic summaries (restricted to 665 bytes) for each cluster. We evaluate our system against the top performing of these 22 systems, according to ROUGE-L, which we denote *top-DUC2004*.[4]

5.1 Automatic Evaluation Using ROUGE

As noted in Section 4.1, we experimented with a number of learning algorithms when building our biographical-sentence classifier. For each machine learning algorithm tested, we build a system that initially classifies the input list of sentences into biographical and non-biographical sentences and then removes redundant sentences. Next, we produce three versions of each system: one which implements a baseline ordering procedure, in which sentences from the clusters are ordered by their appearance in their source document (e.g. any sentence which occurred first in its original document is placed first in the summary, with ties ordered randomly within the set), a second which orders the biographical sentences by the confidence score obtained from the classifier, and a third which uses the SVM regression as the reordering component. Finally, we run our reference rewriting component on each and trim the output to 665 bytes.

We evaluate first using the ROUGE-L metric (Lin and Hovy, 2003) with a 95% (ROUGE computed) confidence interval for all systems and compared these to the ROUGE-L score of the best-performing DUC2004 system.[5] The higher the ROUGE score, the closer the summary is to the DUC2004 human reference summaries. As shown in Figure 1, our best performing system is the multinomial naïve Bayes classifier (MNB) using the classifier confidence scores to order the sentences in the biography. This system significantly outperforms the top ranked DUC2004 system (top-DUC2004).[6] The success of this particularly learning algorithm on our task may be due to: (1) the nature of our feature space – n-gram frequencies are modeled properly by a multinomial distribution; (2) the simplicity of this classifier particularly given our large feature dimensional-

[3] http://duc.nist.gov/duc2004

[4] Note that this system out-performed 19 of the 22 systems on ROUGE-1 and 20 of 22 on ROUGE-L and ROUGE-W-1.2 ($p < .05$) (Blair-Goldensohn et al., 2004a). No ROUGE metric produced scores where this system scored significantly worse than any other system. See Figure 2 below for a comparison of all DUC2004 systems with our top system where all systems are evaluated using ROUGE-L-1.5.5.

[5] We used the same version (1.5.5) of the ROUGE metric to compute scores for the DUC systems and baseline also.

[6] Significance for each pair of systems was determined by paired t-test and calculated at the .05 significance level.

ity; and (3) the robustness of naïve Bayes with respect to noisy data: Not all sentences in Wikipedia biographies are biographical sentences and some sentences in TDT4 *are* biographical.

While the SVM regression reordering component has a slight negative impact on the performance of the MNB system, the difference between the two versions is not significant. Note however, that both the C4.5 and the SVM versions of our system *are* improved by the SVM regression sentence reordering. While neither performs better than top-DUC2004 without this component, the C4.5 system with SVM reordering is significantly better than top-DUC2004 and the performance of the SVM system with SVM regression is comparable to top-DUC2004. In fact, when we use only the SVM regression model to rank the hypothesis sentences, **without** employing any classifier, then remove redundant sentences, rewrite and trim the results, we find that, interestingly, this approach also outperforms top-DUC2004, although the difference is not statistically significant. However, we believe that this is an area worth pursuing in future, with more sophisticated features.

The following biography of Brian Jones was produced by our MNB system and then the sentences were ordered using the SVM regression model:

> Born in Bristol in 1947, Brian Jones, the co-pilot on the Breitling mission, learned to fly at 16, dropping out of school a year later to join the Royal Air Force. After earning his commercial balloon flying license, Jones became a ballooning instructor in 1989 and was certified as an examiner for balloon flight licenses by the British Civil Aviation Authority. He helped organize Breitling's most recent around-the-world attempts, in 1997 and 1998. Jones, 52, replaces fellow British flight engineer Tony Brown. Jones, who is to turn 52 next week, is actually the team's third co-pilot. After 13 years of service, he joined a catering business and, in the 1980s,...

Figure 2 illustrates the performance of our MNB system with classifier confidence score sentence ordering when compared to mean ROUGE-L-1.5.5 scores of DUC2004 human-generated summaries and the 22 DUC2004 systems' summaries across all summary tasks. Human summaries are labeled A-H, DUC2004 systems 1-22, and our MNB system is marked by the rectangle. Results are sorted by mean ROUGE-L score. Note that our system performance is actually comparable in ROUGE-L score to one of the human summary generators and is significantly better that all DUC2004 systems, including top-DUC2004, which is System 1 in the figure.

5.2 Manual Evaluation

ROUGE evaluation is based on n-gram overlap between the automatically produced summary and the human reference summaries. Thus, it is not able to measure how fluent or coherent a summary is. Sentence ordering is one factor in determining fluency and coherence. So, we conducted two experiments to measure these qualities, one comparing our top-performing system according to ROUGE-L score (MNB) vs. the top-performing DUC2004 system (top-DUC2004) and another comparing our top system with two different ordering methods, classifier-based and SVM regression.[7] In each experiment, summaries were trimmed to 665 bytes.

In the first experiment, three native American English speakers were presented with the 50 questions (Who is X?). For each question they were given a pair of summaries (presented in random order): one was the output of our MNB system and the other was the summary produced by the top-DUC2004 system. Subjects were asked to decide which summary was more responsive in form and content to the question or whether both were equally responsive. 85.3% (128/150) of subject judgments preferred one summary over the other. 100/128 (78.1%) of these judgments preferred the summaries produced by our MNB system over those produced by top-DUC2004. If we compute the majority vote, there were 42/50 summaries in which at least two subjects made the same choice. 37/42 (88.1%) of these majority judgments preferred our system's summary (using binomial test, $p = 4.4e-7$). We used the weighted kappa statistic with quadratic weighting (Cohen, 1968) to determine the inter-rater agreement, obtaining a mean pairwise κ of 0.441.

Recall from Section 5.1 that our SVM regression reordering component slightly decreases the average ROUGE score (although not significantly) for our MNB system. For our human evaluations, we decided to evaluate the quality of the presentation of our summaries with and without this compo-

[7] Note that top-DUC2004 was ranked sixth in the DUC 2004 **manual** evaluation, with **no** system performing significantly better for coverage and only 1 system performing significantly better for responsiveness.

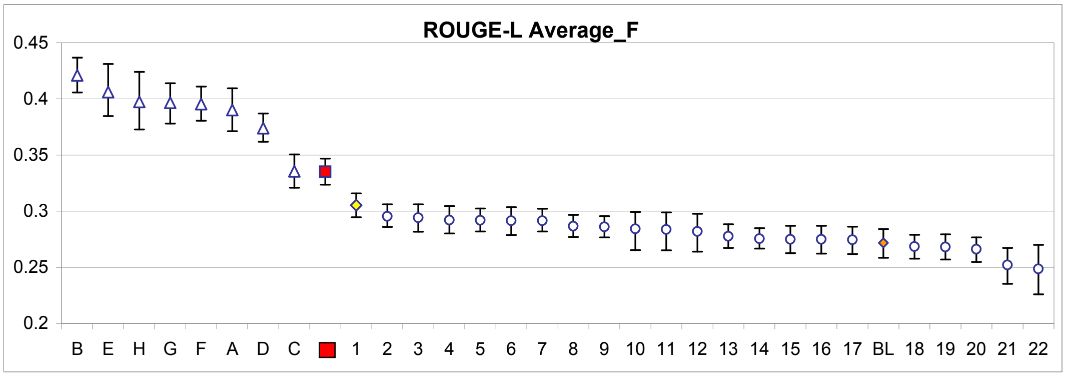

Figure 2: ROUGE-L scores for DUC2004 human summaries (A-H), our MNB system (rectangle), and the DUC2004 competing systems (1-22 anonymized), with the baseline system labeled BL.

nent to see if this reordering component affected human judgments even if it did not improve ROUGE scores. For each question, we produced two summaries from the sentences classified as biographical by the MNB classifier, one ordered by the confidence score obtained by the MNB, in decreasing order, and the other ordered by the SVM regression values, in increasing order. Note that, in three cases, the summary sentences were ordered identically by both procedures, so we used only 47 summaries for this evaluation. Three (different) native American English speakers were presented with the 47 questions for which sentence ordering differed. For each question they were given the two summaries (presented in random order) and asked to determine which biography they preferred.

We found inter-rater agreement for these judgments using Fleiss' kappa (Fleiss, 1971) to be only moderate (κ=0.362). However, when we computed the majority vote for each question, we found that 61.7% (29/47) preferred the SVM regression ordering over the MNB classifier confidence score ordering. Although this difference is not statistically significant, again we find the SVM regression ordering results encouraging enough to motivate our further research on improving such ordering procedures.

6 Related Work

The DUC2004 system achieving the highest overall ROUGE score, our top-DUC2004 in Section 5, was Blair-Goldensohn et al. (2004a)'s DefScriber, which treats "Who is X?" as a definition question and targets definitional themes (e.g. genus-species) found in the input document collections which include references to the target person. Extracted sentences are then rewritten using a reference rewriting system (Nenkova and McKeown, 2003) which attempts to shorten subsequent references to the target. Sentences are ordered in the summary based on a weighted combination of topic centrality, lexical cohesion, and topic coverage scores. A similar approach is explored in Biryukov et al. (2005), which uses Topic Signatures (Lin and Hovy, 2000) constructed around the target individual's name to identify sentences to be included in the biography.

Zhou et al. (2004)'s biography generation system, like ours, trains biographical and non-biographical sentence classifiers to select sentences to be included in the biography. Their system is trained on a hand-annotated corpus of 130 biographies of 12 people, tagged with 9 biographical elements (e.g., bio, education, nationality) and uses binary unigram and bigram lexical and unigram part-of-speech features for classification. Duboue et al. (2003) also address the problem of learning content selection rules for biography. They learn rules from two corpora, a semi-structured corpus with lists of biographical facts about show business celebrities and a corpus of free-text biographies about the same celebrities.

Filatova et al. (2005) learn text features typical of biographical descriptions by deducing biographical and occupation-related activities automatically by compariing descriptions of people with different occupations. Weischedel et al. (2004) models kernel-fact features typical for biographies using linguistic and semantic processing. Linguistic features are derived from predicate-argument structures deduced from parse trees, and semantic features are the set of biography-related relations and events defined in the ACE guidelines (Doddington et al., 2004). Sentences containing kernel facts are ranked using probabilities estimated from a corpus of manually created biographies, including Wikipedia, to estimate the conditional distribution of relevant material given a kernel fact and a background corpus.

The problem of ordering sentences and preserving coherence in MDS is addressed by Barzilay et al. (2001), who combine chronological ordering of events with cohesion metrics. SVM regression has recently been used by (Li et al., 2007) for sentence ranking for general MDS. The authors calculated a similarity score for each sentence to the human summaries and then regress numeric features (e.g., the centroid) from each sentence to this score. Barzilay and Lee (2004) use HMMs to capture topic shift within a particular domain; sequence of topic shifts then guides the subsequent ordering of sentences within the summary.

7 Discussion and Future Work

In this paper, we describe a MDS system for producing biographies, given a target name. We present an unsupervised approach using Wikipedia biography pages and a general news corpus (TDT4) to automatically construct training data for our system. We employ a NE tagger and a coreference resolution system to extract class-based and lexical features from each sentence which we use to train a binary classifier to identify biographical sentences. We also train an SVM regression model to reorder the sentences and then employ a rewriting heuristic to create the final summary.

We compare versions of our system based upon three machine learning algorithms and two sentence reordering strategies plus a baseline. Our best performing system uses the multinomial naïve Bayes (MNB) classifier with classifier confidence score reordering. However, our SVM regression reordering improves summaries produced by the other two classifiers and is preferred by human judges. We compare our MNB system on the DUC2004 biography task (task 5) to other DUC2004 systems and to human-generated summaries. Our system out-performs all DUC2004 systems significantly, according to ROUGE-L-1.5.5. When presented with summaries produced by our system and summaries produced by the best-performing (according to ROUGE scores) of the DUC2004 systems, human judges (majority vote of 3) prefer our system's biographies in 88.1% of cases.

In addition to its high performance, our approach has the following advantages: It employs no manual annotation but relies upon identifying appropriately different corpora to represent our training corpus. It employs class-based as well as lexical features where the classes are obtained automatically from an ACE NE tagger. It utilizes automatic coreference resolution to identify sentences containing references to the target person. Our sentence reordering approaches make use of either classifier confidence scores or ordering learned automatically from the actual ordering of sentences in Wikipedia biographies to determine the order of presentation of sentences in our summaries.

Since our task is to produce concise summaries, one focus of our future research will be to simplify the sentences we extract before classifying them as biographical or non-biographical. This procedure should also help to remove irrelevant information from sentences. Recall that our SVM regression model for sentence ordering was trained using only biographical class-based/lexical items. In future, we would also like to experiment with more linguistically-informed features. While Wikipedia does not enforce any particular ordering of information in biographies, and while different biographies may emphasize different types of information, it would appear that the success of our automatically derived ordering procedures may capture some underlying shared view of how biographies are written. The same underlying views may also apply to domains such as organization descriptions or types of historical events. In future we plan to explore such a generalization of our procedures to such domains.

Acknowledgments

We thank Kathy McKeown, Andrew Rosenberg, Wisam Dakka, and the Speech and NLP groups at Columbia for useful discussions. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001106C0023 (approved for public release, distribution unlimited). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of NAACL-HLT*.

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of the First Human Language Technology Conference*, San Diego, California.

Maria Biryukov, Roxana Angheluta, and Marie-Francine Moens. 2005. Multidocument question answering text summarization using topic signatures. In *Proceedings of the 5th Dutch-Belgium Information Retrieval Workshop*, Utrecht, the Netherlands.

Sasha Blair-Goldensohn, David Evans, Vasileios Hatzivassiloglou, Kathleen McKeown, Ani Nenkova, Rebecca Passonneau, Barry Schiffman, Andrew Schlaikjer, Advaith Siddharthan, and Sergey Siegelman. 2004a. Columbia University at DUC 2004. In *Proceedings of the 4th Document Understanding Conference*, Boston, Massachusetts, USA.

Sasha Blair-Goldensohn, Kathy McKeown, and Andrew Schlaikjer. 2004b. Answering definitional questions: A hybrid approach. In Mark Maybury, editor, *New Directions In Question Answering*, chapter 4. AAAI Press.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. volume 70, pages 213–220.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction program - tasks, data, and evaluation. In *Proceedings of the LREC Conference*, Canary Islands, Spain, July.

Pablo Duboue and Kathleen McKeown. 2003. Statistical acquisition of content selection rules for natural language generation. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing*, pages 121–128, Sapporo, Japan, July.

Elena Filatova and John Prager. 2005. Tell me what you do and I'll tell you what you are: Learning occupation-related activities for biographies. In *Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 113–120, Vancouver, Canada, October.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. volume 76, No. 5, pages 378–382.

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. Nyu's english ace 2005 system description. In *ACE 05 Evaluation Workshop*, Gaithersburg, MD.

Sujian Li, You Ouyang, Wei Wang, and Bin Sun. 2007. Multi-document summarization using support vector regression. In *http://duc.nist.gov/pubs/2007papers*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 495–501, Saarbrücken, Germany, July.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Language Technology Conference*, Edmonton, Canada.

Ani Nenkova and Kathleen McKeown. 2003. References to named entities: A corpus study. In *Proceedings of the Joint Human Language Technology Conference and North American chapter of the Association for Computational Linguistics Annual Meeting*, Edmonton, Canada, May.

Ralph Weischedel, Jinxi Xu, and Ana Licuanan. 2004. A hybrid approach to answering biographical questions. In Mark Maybury, editor, *New Directions In Question Answering*, chapter 5. AAAI Press.

I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham. 1999. Weka: Practical machine learning tools and techniques with java implementation. In *International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems*, pages 192–196.

Liang Zhou, Miruna Ticrea, and Eduard Hovy. 2004. Multi-document biography summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 434–441, Barcelona, Spain.