SUMMARIZATION- AND LEARNING-BASED APPROACHES TO INFORMATION DISTILLATION

Boriska Toth[1,2]*, Dilek Hakkani-Tür*[2]*, Sibel Yaman*[2]

[2]University of California, Berkeley, CA,
[2]International Computer Science Institute, Berkeley, CA
bori@eecs.berkeley.edu, {dilek,sibel}@icsi.berkeley.edu

ABSTRACT

Information distillation is the task that aims to extract relevant passages of text from massive volumes of textual and audio sources, given a query. In this paper, we investigate two perspectives that use shallow language processing for answering open-ended distillation queries, such as "List me facts about [event]". The first approach is a summarization-based approach that uses the unsupervised maximum marginal relevance (MMR) technique to successfully capture relevant but not redundant information. The second approach is based on supervised classification and trains support vector machines (SVMs) to discriminate relevant snippets from irrelevant snippets using a variety of features. Furthermore, we investigate the merit of using the ROUGE metric for its ability to evaluate redundancy alongside the conventionally used F-measure for evaluating distillation systems. Our experimental results with textual data indicate that SVM and MMR perform similarly in terms of ROUGE-2 scores while SVM is better than MMR in terms of F1 measure. Moreover, when speech recognizer output is used, SVM outperforms MMR in terms of both scores.

Index Terms— information distillation, information extraction, summarization, supervised learning, speech processing

1. INTRODUCTION

According to its description in the context of DARPA-funded GALE program, an information distillation system is given a query that conforms to one of the question templates that are in the form of either broad questions or questions focused on a particular information need. A large corpus (typically in the thousands range) of textual and audio documents is also provided from which relevant sentences are to be extracted. Information distillation is in fact a specific variant of the fundamental problems of information extraction and automated question answering in which the goal is to return all those sentences that a human judge would find relevant to the given query. This is a challenging task since identifying relevant sentences is a highly subjective task even for humans. Furthermore, relevance judgments of sentences are usually interdependent as the meaning of some sentences arises only from context.

In our previous work, we presented a *generic* data-driven method for sentence extraction using lexical and name matching features for various query templates with different forms [1]. We later extended this work to also include syntactic, semantic, information extraction (IE) annotation, and topicality features [2]. In this paper, we specifically focus on open-ended descriptive queries, in particular those that come from Template 1 of the DARPA GALE program with the form *Describe facts about [event]*. An example query is *List facts about [Hillary Clinton's election to the US Senate]*. This template is especially difficult to answer with the learning-based approach as it is highly open-ended. Contrary to other templates, its form is very general, making it difficult to find trainable patterns for this template. For this reason, specific approaches have been proposed for answering queries coming from Template 1. In [3], a textual entailment-based approach using syntactic and semantic parsing was proposed. The slot and the sentences are represented as a number of "proposition trees", which is a hierarchy of interconnected elementary predicate-argument structures, and slot trees are instantiated in the sentence trees. In [4], an iterative unsupervised sentence extraction method that finds the subset of sentences that are very likely to be relevant or irrelevant and iteratively trains a classification model using these examples was proposed.

In many respects, answering Template 1 queries is similar to the query-based summarization of Document Understanding Conference (DUC) evaluations [5]. In fact, summarization-based methods for distillation have been investigated in previous work. In [6], a comparison is made between a query-based summarization approach and a statistical information retrieval and extraction system. They consider the "prosecution" template (seeking answers to queries of the form "Describe the prosecution of [person] for [crime]") that may favor learning because of availability of trainable patterns such as "[person] is accused of". While they report results in favor of the learning-based system, they make use of complex features in its design. In [7], authors study the task of producing biographies which typically have a certain structure, e.g., the birth date comes before the education information. They take a query-based summarization approach and employ supervised learning to filter and reorder the extracted sentences. It is important to keep in mind that all these previous works have focused on text data, many times well-structured data which makes it possible to derive highly accurate rule-based systems. With the exception of the work in [1], it still remains an unexplored question to study the effects of using automatic speech recognition (ASR) transcriptions to training and/or test a distillation system. It is crucial to build information distillation systems robust to ASR errors to make sensible use of audio resources.

In this paper, we investigate the use of a summarization-based approach in comparison to a learning-based approach for Template 1 as described above. As for the first approach, the maximum marginal relevance (MMR) summarization technique [8] aims to capture relevant but not redundant information. This goal is achieved by scoring sentences according to their similarity to the given query while penalizing their similarity to the already selected sentences. As the second approach that we investigate in this paper, supervised learning-based approach benefits from the available annotated data to discriminate relevant snippets from irrelevant snippets that leverage a variety of query-independent and query-dependent features, including word n-grams and similarity of a candidate sentence to

	Textual	Speech
No. Queries	49	10
No. Documents	913	81
No. Relevant Sentences	2,854	250
No. Irrelevant Sentences	15,874	1,581

Table 1. Statistics of the datasets we use.

the given query. In addition, since the performance of summarization systems is evaluated in terms of ROUGE scores, we investigate the merit of using the ROUGE metric in the distillation task for its ability to evaluate redundancy alongside the conventionally used F1-measure. Our motivation for this investigation comes from the fact that F1-measure underestimates the performance of summarization methods as it does not punish redundancy. As another contribution, we study and quantify the effects of training and testing on speech data versus textual data. Our experimental results with textual data indicate that SVM and MMR perform similarly in terms of ROUGE-2 scores while SVM is better than MMR in terms of F1 measure. Moreover, when ASR transcriptions are used, SVM outperforms MMR in terms of both scores. As further experiments demonstrate, an advantage of MMR is its robustness to speech data compared to SVM when SVM is trained on textual data.

2. ANSWERING OPEN-ENDED DESCRIPTIVE QUERIES

2.1. Datasets and Feature Extraction

We work with three datasets from the GALE distribution: Textual, Speech-ASR, and Speech-Manual. *Textual* refers to documents from textual sources, such as newswire, *Speech-ASR* documents are ASR transcriptions of audio resources obtained using SRI's speech recognizer, and *Speech-Manual* are the corresponding manual transcriptions. The word error rate (WER) of the speech data is 28.2% when compared against the closed-caption manual transcriptions.The statistics of these three datasets are shown in Table 1.

We performed simple text cleaning with simple pattern matching rules. Our experiments with stemming and stop-word removal did not yield noticeable improvements. We padded the query text with given related terms (if any), which are typically synonyms or closely related terms to the term in the query. We use lexical features consisting of unigrams and bigrams. We experimented with extracting three kinds of features: indicators, term frequency (fraction of times a term appears in an example), and TF-IDF (term frequency multiplied by inverse document frequency). We found TF-IDF to outperform the other features, across both algorithms and all datasets, and hence we use TF-IDF features in all our experiments.

2.2. Summarization-Based Distillation Using MMR

Maximum marginal relevance (MMR) is a very simple, yet highly effective, algorithm for summarization. At each step, it maintains text Sum containing the sentences extracted into the summary so far, and a list of sentences, $\{S_i\}$, yet to be extracted. In each iteration, the top-scoring sentence is added to the summary. Each sentence, represented with unigram and bigram TF-IDF values, is assigned a relevancy score by weighting its similarity to the query and penalizing the redundancy with the summary-so-far

$$MMR(S_i) = \lambda \cdot Sim(S_i, Query) - (1-\lambda) \cdot Sim(S_i, Sum), \quad (1)$$

where $0 < \lambda < 1$ is a weighting parameter, and Sim is the cosine similarity between two feature vectors. The optimal value of λ is found so that F1 measure is optimized on a cross-validation set.

An inspection of the answer keys for Template 1 reveals that redundancy may be an issue as, for instance, many times the answering sentences deal not only with the specific details of how an event happened but also with other aspects of the events such as the statements from officials regarding the event and the public reactions to the event. Furthermore, the relevancy of these latter type of sentences is somewhat loosely defined and can be based on the fact that they introduce a novel aspect that is not mentioned in another sentence.

MMR's major strength in answering open-ended descriptive queries such as those from Template 1 is its capability to promote sentences with novel information, as evidenced by Eq. 1. The first term, $Sim(S_i, Query)$ stands for how similar the given sentence is to the given query. Once the most relevant sentence is selected and added to the summary, what we would like to add next is a sentence that is still highly relevant to the query yet introduces some novel information. In this way redundancy is reduced so as to discover other facets of the problem, i.e., rather than repeating the material that is already mentioned in the summary, a different aspect is introduced. This is achieved by the second term, which is the similarity between S_i and the summary composed so far, Sum. By iterating this selection process, one continues adding more sentences until the percentage of the sentences that optimized a desired performance (e.g., F1 value) on a development set are selected.

2.3. Learning-Based Distillation Using SVM

The distillation problem can also be posed as a binary classification task where the goal is to discriminate relevant from irrelevant sentences. A supervised training approach makes it possible to take advantage of available annotated data to generalize on what should typically be included in the answer and to incorporate a variety of features that capture many dimensions of the problem. In the training stage, one can assign a class label to every candidate sentence by labeling the sentences in the manually-prepared answer keys as relevant and the rest of the sentences in the relevant documents as irrelevant.

For supervised learning, we performed preliminary experiments using both support vector machines (SVMs) [9, 10] and AdaBoost [11], and found SVM to generally outperform boosting. In this work, we use as features the cosine similarity between unigram and bigram lexical feature vectors of the query and the sentence, TF-IDF-weighted unigram and bigram lexical features, and the position as a fraction of the sentences in a document[1].

2.4. Evaluation of Distillation Systems

F1 measure is the commonly used scoring metric for the distillation task. It is the natural metric to use when taking the perspective of distillation as the problem of choosing to include or not include individual sentences. In this work, we explore the use of ROUGE-2, which is the bigram overlap of text output by a distillation system to the reference text of an answer key. These two evaluation measures do not necessarily change in parallel.

While the recall-oriented ROUGE score[12] is the metric of choice for evaluation of summarization systems, its use has typically been overlooked for distillation. One basic argument for using a summarization-oriented metric in scoring distillation is that the selection of relevant sentences for the answer keys is observed to be highly subjective. Thus, it can be valuable to use a metric that scores the overall information content, as opposed to the correctness in selecting particular sentences.

A second reason to incorporate ROUGE scoring is that a classification method is intuitively better tailored to optimizing F-measure than summarization is, while a summarization approach like MMR

[1] For instance, a sentence at the beginning (end) of a document has a feature value of 0. The motivation for these features is that a sentence at the beginning of a document is likely to be relevant.

Algorithm	Train set	F-measure	Improvement over baseline
Baseline	-	.290	0
MMR	-	.343	.053
SVM	Textual	.369	.079

Table 2. Comparison of summarization and learning for Textual dataset. In the baseline system, all examples are labeled as positive.

is conceptually targeted to optimizing the overall information content. At the same time, the redundancy penalization in MMR does not favor a per-sentence scoring scheme. ROUGE directly handles redundancy issue by scoring multiple appearances only once, i.e., if a sentence appears twice in the answer key, ROUGE ignores the duplicate while F1 does not.

A third merit of using ROUGE is to score the quality of distillation outputs that use speech recognition data. F-measure fails to capture whether the extracted sentences have high fidelity to the spoken speech they represent, or whether ASR errors greatly degrade the information content or intelligibility of the extracted text.

3. EXPERIMENTS

We conducted experiments to investigate the performance of the summarization-based and learning-based approaches using text and audio data. We explored the role of the number of sentences included in the final answer. We also investigated the appropriateness of F1 and ROUGE scores in this task.

3.1. Experimental Setup

A key motivation in our experimental setup was to remove variability in our results that comes about from the choice of development, training, and test sets. We achieve this goal through averaging large numbers of runs of each experiments, where randomization is used in each run to split the data into development, training (for the case of SVM), and test sets.

For MMR, for each run of an experiment, we split the dataset into development and test sets by choosing queries uniformly at random for the development set, until at least 800 examples have been chosen for the development set. We repeat the above described experiment $n = 10$ times for the case of using the Textual dataset (which is quite large), and $n = 100$ times for the case of using a Speech dataset, and report averaged results. The two development parameters we optimize are λ (the tradeoff between similarity to the query vs the redundancy penalization), and the percent r of sentences chosen for an answer.

For classification, we use a variant of n-fold cross validation, in which we use each query in a dataset as the test query once, and randomly split the rest of the data into training and development sets. We macro-average results across all test queries. We perform n=10 trials of the n-fold cross validation procedure for Textual dataset and n=50 trials for the Speech datasets. The development parameters here are a regularization parameter for the SVM optimization problem, and the percent of sentences r to select.

With these choices of n, our averaged results are in practice invariable up to 4 digits in both cases. Note that averaged results include at least 400 instances of running a randomized experiment on a test query in all cases of what algorithm and dataset we use.

3.2. Evaluation in terms of F1-measure

We first analyze the performance of our algorithms in terms of F-measure. In interpreting the F-measure given here, one must keep in mind that, for the Template 1 distillation task, improvements in F-measure over the trivial baseline reported in the literature are generally less than 10% [4].

	Test set	
	Speech-Man	Speech-ASR
Baseline	0.335	0.335
MMR	0.384	0.376
SVM - Textual train set	0.418	0.395
SVM - Speech-Manual train set	0.396	0.374
SVM - Speech-ASR train set	-	0.428

Table 3. Testing on Speech-Manual and Speech-ASR data, using both summarization and learning with various train sets.

Table 2 compares learning and summarization on the large Textual dataset. The summarization approach achieves a 5.3% improvement over the baseline (perfect recall, low precision), while learning achieves 7.9%. The corresponding numbers for Speech-ASR data, in table 3, are 4.1% improvement for summarization, and 9.3% improvement for learning in the best case. While supervised learning outperforms summarization in both cases, MMR achieves much of the improvement that SVM does. Further, unlike supervised learning, the summarization approach does not require manually-annotated training data or the complexity of tuning features.

In table 3, we summarize results for working with speech data, which in this case is very noisy (WER of 16%). In the case of using SVM and testing on Speech-ASR, using Speech-ASR data as the training set proved the most effective ($F_1 = .428$), even though the training set is about one-fifteenth the size as when the Textual dataset is used for training. Further, we can quantify the degradation in performance due to speech recognition noise by comparing the performance of Speech-ASR vs Speech-Manual test sets when the setup is the same. For MMR, the drop when testing on Speech-ASR data as compared to Speech-Manual is .8%; for SVM trained on Textual, it's 2.3%; and for SVM trained on Speech-Manual it's 2.2%. MMR is more robust to ASR noise than SVM trained on textual or manually transcribed data. This makes sense given that MMR isn't trained and works with the characteristics of the ASR test data, while SVM is classifying based on training on clean, non-ASR data.

Finally, we point out that the F-measure achieved (.428) on Speech-ASR data trained on Speech-ASR data is competitive with some results cited in the literature for textual data, despite having a very high WER and the small size of the training set [4]. This shows the effectiveness of training on ASR-generated data.

3.3. Evaluation in terms of ROUGE

In this section, we interpret distillation results with the ROUGE scores. Since ROUGE is recall-oriented, we report it as a function of the length of the returned answer. Sentences are sorted by the algorithm used (using classification scores in the case of SVM and the selection process in the case of MMR), and the top r% of either words or sentences is returned. Each plot of rouge curves we show is the interpolation of the average of at least 400 instances of testing on a query at each point.

Figure 1 shows how ROUGE-2 score degrades for speech recognition as compared to manually transcribed sentences. The top curve gives ROUGE as a function of the percent of sentences chosen using Speech-Manual data. The other 3 curves use Speech-ASR data for testing. Curve (3) isolates the effect of ASR noise: sentences are ordered in the exact ordering as for curve (1), and in going from curve (1) to (3), each Speech-Manual sentence is replaced by its corresponding Speech-ASR sentence. Thus, all difference in Rouge-2 scores is due to ASR noise. The difference in Rouge is about 7% at the percent of sentences chosen $r = 30\%$ point that we operate near [2]. In curve (4), the Speech-ASR sentences are sorted by the

[2]The percent of sentences to choose for optimal F-measure was found during parameter tuning to be 30%, and percent of words to choose 35%.

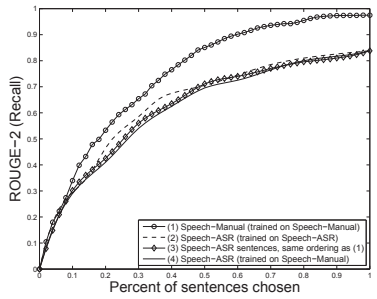

Fig. 1. Rouge-2 as a function of percent of sentences chosen for the answer

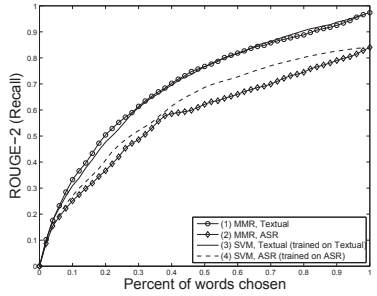

Fig. 2. Rouge-2 as a function of percent of words chosen for the answer

classification scores generated by SVM trained on Speech-Manual. This curve represents slightly lower Rouge scores than (3). Finally, curve (2) shows that if Speech-ASR data is used to train an SVM, the drop in Rouge scores due to ASR noise is mitigated somewhat. From Figure 1 we conclude that the drop in Rouge-2 due to speech recognition errors is substantial: when all example sentences are output in the answer generated by the system, roughly 17% of bigrams appearing in the answer key do not appear due to ASR transcription errors (incidentally, this value is quite close to the WER).

We next compare Rouge-2 scores arising in the summarization vs learning cases. Figures 2 and 3 give Rouge as a function of percent of words chosen, and percent of sentences chosen, respectively. For Textual data, the two algorithms perform quite similarly: using percent of words chosen, the two curves are almost indistinguishable. Testing on Speech-ASR data, on the other hand, SVM (trained on Speech-ASR) beats MMR in Rouge-2 along the entire curve in figure 2. However, much of this difference between SVM and MMR on Speech-ASR data vanishes when giving Rouge as a function of percent of sentences (see figure 3). This means MMR tends to pick longer sentences relative to SVM. The summarization and learning-based approaches in fact have very similar Rouge-performance at the level of sentence or word selection they operate at ($30\% \leq r \leq 40\%$): performance is virtually identical in this region of the x-axis for Textual data, and Rouge-2 scores are within 2% in this region for Speech-ASR data.

We intuitively expected summarization (MMR) to outperform learning (SVM) in Rouge score because MMR penalizes redundancy, which might make for higher bigram overlap to the reference answer. To justify why this did not happen, we note that SVM outperforms MMR in sentence-level recall in all the comparisons we made in this section. Thus, SVM picks more relevant sentences than MMR at a fixed value for the number of sentences returned, which can reverse the effect that MMR might be better at picking sentences with high information content among the relevant sentences.

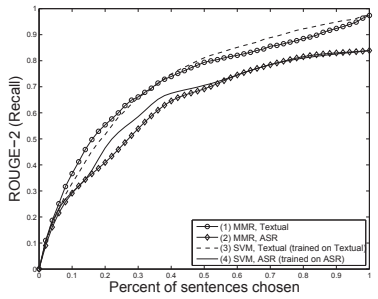

Fig. 3. Rouge-2 as a function of percent of sentences chosen for the answer

4. CONCLUSIONS AND FUTURE WORK

This work presents two approaches to the information distillation task: summarization using MMR and supervised learning using SVM. While SVM outperforms MMR for both Textual and ASR-generated data, MMR has the advantage of not requiring human-annotated training data. MMR achieves much of the improvement in F-measure as SVM for both Textual and Speech data, and it displays similar performance in Rouge-2 score (within 2%). We also demonstrate the value of Rouge score for evaluating distillation systems, for instance to measure the quality of ASR-generated noisy answers.

In our future work, we plan to incorporate confidence values generated by the ASR system for ASR-output words as a learning feature, as well as to investigate other ways to enhance the robustness of distillation systems to ASR-generated noise.

5. REFERENCES

[1] D. Hakkani-Tür and G. Tür, "Statistical sentence extraction for information distillation," in *Proc. of ICASSP*, Honolulu, HI, USA, 2007.

[2] M. Levit, D. Hakkani-Tür, G. Tür, and D. Gillick, "Ixir: A statistical information distillation system," *Computer Speech and Language*, vol. 23, no. 4, pp. 527–542, 2009.

[3] M. Levit, E. Boschee, and M. Freedman, "Selecting On-Topic Sentences from Natural Language Corpora," in *Proceedings of Interspeech-2007*, Antwerp, Belgium, August 2007b, pp. 2793–2796.

[4] K. Kamangar, D. Hakkani-Tür, G. Tür, and M. Levit, "An iterative unsupervised learning method for information distillation," *IEEE ICASSP, Las Vegas, NV*, 2008.

[5] K. McKeown, "Lessons learned from large scale evaluation of systems that produce text: nightmares and pleasant surprises," in *NAACL '03: Proceedings of the Fourth International Natural Language Generation Conference*, Sydney, July 2006, pp. 3–5.

[6] B. Schiffman, K. McKeown, R. Grishman, and J. Allan, "Question answering using integrated information retrieval and information extraction," in *HLT-NAACL*, 2007, pp. 532–539.

[7] F. Biadsy, J. Hirschberg, and E. Filatova, "An unsupervised approach to biography production using wikipedia," in *Proceedings of ACL-HLT*, Columbus, Ohio, June 2008, pp. 807–815.

[8] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Research and Development in Information Retrieval*, 1998, pp. 335–336.

[9] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. of the European Conference on Machine Learning (ECML)*, Berlin, 1998, pp. 137–142.

[10] T. Joachims, "Making large-scale svm learning practical," LS8-Report 24, Universität Dortmund, LS VIII-Report, 1998.

[11] Y. Freund and R. E. Schapire, "A short introduction to boosting," in *Proc. of the International Joint Conference on Artificial Intelligence*, 1999, pp. 1401–1406.

[12] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *NAACL-HLT*, Morristown, NJ, USA, 2003, pp. 71–78.