Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization

Proceedings of the ACL-05 Workshop

Organizing Committee
Jade Goldstein, US Department of Defense
Alon Lavie, CMU Language Technologies Institute
Chin-Yew Lin, USC Information Sciences Institute
Clare Voss, US Army Research Laboratory

29 June 2005
University of Michigan
Ann Arbor, Michigan, USA

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Preface

This workshop is the first meeting to focus on the challenges that the machine translation (MT) and summarization communities face in developing valid and useful evaluation measures. Our aim is to bring these two communities together to learn from each other's approaches.

Prior ACL workshops on evaluation have had as their central focus a core computational task (e.g., word sense disambiguation, parsing), a genre (e.g., dialogue, multi-modal interfaces), a computational technique (e.g., unsupervised learning, finite state models), a resource (e.g., parallel texts, WordNet), or a process (e.g., reading comprehension, question-answering). This workshop, in clear contrast, has as its central focus the examination of evaluation measures, or "meta-evaluation" as Dan Melamed has noted.

The initial impetus for this workshop came at the biennial meeting of the Association for Machine Translation in the Americas (AMTA) held at Georgetown University in September 2004, when the following question arose in a discussion session: "Why isn't recall a part of MT evaluation the way that it is for summarization evaluation?" Several of us continued this discussion afterwards and proposed to convene together again more formally to address this question and other evaluation challenges that both the MT and summarization communities have been tackling.

We wish to thank Bonnie Dorr and Ed Hovy, in particular, for their encouragement and contributions in shaping the initial workshop proposal and the subsequent call for papers. Boyan Onyshkevych, Barb Wheatley, Donna Harmon, and Judith Klavans also provided insightful comments in the proposal writing phase of the workshop that helped guide and focus the topics we chose to address.

We also would like also to thank several others. Charles Wayne, Joe Olive, Donna Harmon, Hoa Dang, Lori Buckland, and Chris Cieri were critical in helping make the datasets available to workshop participants. Jason Eisner and Philipp Koehn, ACL publications chairs, provided us invaluable assistance in preparing the proceedings.

Many thanks also to the Program Committee and additional reviewers who graciously spent time with a short schedule to review submitted papers and provide valuable feedback. We have an exciting program for which we thank the many authors who submitted their research papers.

Jade Goldstein, Alon Lavie, Chin-Yew Lin, Clare Voss
June 2005

Excerpts from Call for Papers

This one-day workshop will focus on the challenges that the MT and summarization communities face in developing valid and useful evaluation measures. Our aim is to bring these two communities together to learn from each other's approaches.

In the past few years, we have witnessed—in both MT and summarization evaluation—the innovation of ngram-based intrinsic metrics that automatically score system-outputs against human-produced reference documents (e.g., IBM's BLEU and ISI/USC's counterpart ROUGE). Similarly, there has been renewed interest in user applications and task-based extrinsic measures in both communities (e.g., DUC'05 and TIDES'04). Most recently, evaluation efforts have tested for correlations to cross-validate independently derived intrinsic and extrinsic assessments of system-outputs with each other and with human judgments on output, such as accuracy and fluency.

The concrete questions that we hope to see addressed in this workshop include, but are not limited to:

- How adequately do intrinsic measures capture the variation between system-outputs and human-generated reference documents (summaries or translations)? What methods exist for calibrating and controlling the variation in linguistic complexity and content differences in input test-sets and reference sets? How much variation exists within these constructed sets? How does that variation affect different intrinsic measures? How many reference documents are needed for effective scoring?
- How can intrinsic measures go beyond simple n-gram matching, to quantify the similarity between system-output and human-references? What other features and weighting alternatives lead to better metrics for both MT and summarization? How can intrinsic measures capture fluency and adequacy? Which types of new intrinsic metrics are needed to adequately evaluate non-extractive summaries and paraphrasing (e.g.,interlingual) translations?
- How effectively do extrinsic (or proxy extrinsic) measures capture the quality of system output, as needed for downstream use in human tasks, such as triage (document relevance judgments), extraction (factual question answering), and report writing; and in automated tasks, such as filtering, information extraction, and question-answering? For example, when is an MT system good enough that a summarization system benefits from the additional information available in the MT output?
- How should metrics for MT and summarization be assessed and compared? What characteristics should a good metric possess? When is one evaluation method better than another? What are the most effective ways of assessing the correlation testing and statistical modeling that seek to predict human task performance or human notions of output quality (e.g., fluency and adequacy) from "cheaper" automatic metrics? How reliable are human judgments?

Anyone with an interest in MT or summarization evaluation research or in issues pertaining to the combination of MT and summarization is encouraged to participate in the workshop. We are looking for research papers on the aforementioned topics, as well as position papers that identify limitations in current approaches and describe promising future research directions.

To faciliate the comparison of different measures during the workshop, we will be making available data sets in advance for workshop participants to test their approaches to evaluation. Although the shared data sets are separated, we would encourage participants to apply their automatic metrics on both data sets and report comparative results in the workshop.

Shared Data Sets

Shared Data Set for MT Evaluation:

The shared data set consists of the 2003 TIDES MT-Eval Test Data for both Chinese-to-English and Arabic-to-English MT. For each of these two language-pair data sets, the following is provided:

- The set of test sentences in the original source language (Chinese or Arabic)
- MT system output for the set of sentences for 7 different MT systems
- A collection of 4 reference translations (human translated) into English
- Human judgments of MT quality (adequacy and fluency) for the various MT system translations of every sentence. Each sentence was judged by two subjects, each of which assigned both an adequacy score and a fluency score, in the integer range of [1-5].

Shared Data Set for Summarization Evaluation:

The summarization shared data set consists of four years' worth of data from past Document Understanding Conferences (DUC) including:

- Documents
- Summaries, results, etc.
  - Manually created summaries
  - Automatically created baseline summaries
  - Submitted summaries created by the participating groups' systems
  - Tables with the evaluation results
  - Additional supporting data and software

Organizers:

Jade Goldstein	US Department of Defense, USA
Alon Lavie	CMU Language Technologies Institute, USA
Chin-Yew Lin	USC Information Sciences Institute, USA
Clare Voss	Army Research Laboratory, USA

Program Committee:

Yasuhiro Akiba	ATR, Japan
Leslie Barrett	TransClick, USA
Bonnie Dorr	University of Maryland, USA
Tony Hartley	University of Leeds, UK
John Henderson	MITRE, USA
Chiori Hori	CMU Language Technologies Institue, USA
Eduard Hovy	USC Information Sciences Insititute, USA
Doug Jones	MIT Lincoln Laboratory, USA
Philipp Koehn	University of Edinburgh, UK
Marie-Francine Moens	Katholieke Universiteit, Leuven, Belgium
Hermann Ney	RWTH Aachen, Germany
Franz Och	Google, USA
Rebecca Passonneau	Columbia University, USA
Andrei Popescu-Belis	University of Geneva ISSCO/TIM/ETI, Switzerland
Dragomir Radev	University Michigan, USA
Karen Sparck Jones	University of Cambridge Computer Laboratory, UK
Simone Teufel	University of Cambridge Computer Laboratory, UK
Nicola Ueffing	RWTH Aachen, Germany
Hans van Halteren	University of Nijmegen, The Netherlands
Michelle Vanni	Army Research Laboratory, USA
Dekai Wu	HKUST, Hong Kong

Additional Reviewers:

Chad Langley	US Department of Defense, USA
Gregor Leusch	RWTH Aachen, Germany
Klaus Macherey	Google,USA
Wolfang Macherey	Google,USA
Christof Monz	University of Maryland, USA
Judith Schlesinger	IDA Center for Computing Sciences, USA

Table of Contents

Workshop Program

Wednesday, June 29, 2005

8:45–8:50 Opening Remarks

Session 1: Summarization Metrics I

8:50–9:15 *A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?*
Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz and David Zajic

9:15–9:40 *On the Subjectivity of Human Authored Summaries*
BalaKrishna Kolluru and Yoshihiko Gotoh

Session 2: MT Metrics I

9:40–10:05 *Preprocessing and Normalization for Automatic Evaluation of Machine Translation*
Gregor Leusch, Nicola Ueffing, David Vilar and Hermann Ney

10:05–10:30 *Syntactic Features for Evaluation of Machine Translation*
Ding Liu and Daniel Gildea

10:30–11:00 Break

Session 3: Invited Talk

11:00–12:00 *Results of the Multilingual Summarization Evaluation*, Kathy McKeown

Session 4: Student Session - Work in Progress

12:00–12:15 *Evaluation of Sentence Selection on Spoken Dialogue*, Xiaodan Zhu

12:15–12:30 *Toward a Predictive Statistical Model of Task-based Performance*, Calandra R. Tate

12:30–2:15 Lunch

Session 5: Summarization Metrics II

2:15–2:40 *Evaluating Automatic Summaries of Meeting Recordings*
Gabriel Murray, Steve Renals, Jean Carletta and Johanna Moore

2:40–3:05 *Evaluating Summaries and Answers: Two Sides of the Same Coin?*
Jimmy Lin and Dina Demner-Fushman

3:05–3:30 *Evaluating DUC 2004 Tasks with the QARLA Framework*
Enrique Amigó, Julio Gonzalo, Anselmo Peñas and Felisa Verdejo

Session 6: MT Metrics II

4:00–4:25 *On Some Pitfalls in Automatic Evaluation and Significance Testing for MT*
Stefan Riezler and John T. Maxwell

4:25–4:50 *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*
Satanjeev Banerjee and Alon Lavie

Session 7: Panel Discussion and Open Forum on Future Plans

4:50–5:50 Panel Discussion

5:50–6:00 Future Plans

A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?

Bonnie J. Dorr and **Christof Monz** and **Stacy President** and **Richard Schwartz**[†] and **David Zajic**

Department of Computer Science and UMIACS
University of Maryland
College Park, MD 20742
{bonnie,christof,stacypre,dmzajic}@umiacs.umd.edu

[†]BBN Technologies
9861 Broken Land Parkway
Columbia, Maryland 21046
schwartz@bbn.com

Abstract

This paper demonstrates the usefulness of summaries in an extrinsic task of relevance judgment based on a new method for measuring agreement, *Relevance-Prediction*, which compares subjects' judgments on summaries with their own judgments on full text documents. We demonstrate that, because this measure is more reliable than previous gold-standard measures, we are able to make stronger statistical statements about the benefits of summarization. We found positive correlations between ROUGE scores and two different summary types, where only weak or negative correlations were found using other agreement measures. However, we show that ROUGE may be sensitive to the choice of summarization style. We discuss the importance of these results and the implications for future summarization evaluations.

1 Introduction

People often prefer to read a summary of a text document, e.g., news headlines, scientific abstracts, movie previews and reviews, and meeting minutes. Correspondingly, the explosion of online textual material has prompted advanced research in document summarization. Although researchers have demonstrated that users can read summaries faster than full text (Mani et al., 2002) with some loss of accuracy, researchers have found it difficult to draw strong conclusions about the usefulness of summarization due to the low level of interannotator agreement in the gold standards that they have used. Definitive conclusions about the usefulness of summaries would provide justification for continued research and development of new summarization methods.

To investigate the question of whether text summarization is useful in an extrinsic task, we examined human performance in a relevance assessment task using a human text *surrogate* (i.e. text intended to stand in the place of a document). We use single-document English summaries as these are sufficient for investigating task-based usefulness, although more elaborate surrogates are possible, e.g., those that span more than one document (Radev and McKeown, 1998; Mani and Bloedorn, 1998).

The next section motivates the need for developing a new framework for measuring task-based usefulness. Section 3 presents a novel extrinsic measure called *Relevance-Prediction*. Section 4 demonstrates that this is a more reliable measure than that of previous gold standard methods, e.g., the *LDC-Agreement* method used for SUMMAC-style evaluations, and that this reliability allows us to make stronger statistical statements about the benefits of summarization. We expect these findings to be important for future summarization evaluations.

Section 5 presents the results of correlation between task usefulness and the Recall Oriented Understudy for Gisting Evaluation (ROUGE) metric (Lin and Hovy, 2003).[1] While we show that ROUGE correlates with task usefulness (using our Relevance-Prediction measure), we detect a slight difference between informative, *extractive* headlines (containing words from the full document) and less informative, *non-extractive* "eye-catchers" (containing words that might not appear in the full document, and intended to entice a reader to read the entire document).

Section 6 further highlights the importance of this point and discusses the implications for automatic evaluation of non-extractive summaries. To evaluate non-extractive summaries reliably, an automatic measure may require knowledge of sophisticated meaning units.[2] It is our hope that the conclusions drawn herein will prompt investigation into more sophisticated automatic metrics as researchers shift their focus to non-extractive summaries.

[1] ROUGE has been previously used as the primary automatic evaluation metric by NIST in the 2003 and 2004 DUC Evaluations.

[2] The *content units* proposed in recent methods (Nenkova and Passonneau, 2004) are a first step in this direction.

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 1–8, Ann Arbor, June 2005.

2 Background

In the past, assessments of usefulness involved a wide range of both intrinsic and extrinsic (task-based) measures (Sparck-Jones and Gallier, 1996). Intrinsic evaluations focus on coherence and informativeness (Jing et al., 1998) and often involve quality comparisons between automatic summaries and reference summaries that are pre-determined to be of high quality. Human intrinsic measures determine quality by assessing document accuracy, fluency, and clarity. Automatic intrinsic measures such as ROUGE use n-gram scoring to produce rankings of summarization methods.

Extrinsic evaluations concentrate on the use of summaries in a specific task, e.g., executing instructions, information retrieval, question answering, and relevance assessments (Mani, 2001). In relevance assessments, a user reads a topic or event description and judges relevance of a document to the topic/event based solely on its summary.[3] These have been used in many large-scale extrinsic evaluations, e.g., SUMMAC (Mani et al., 2002) and the Document Understanding Conference (DUC) (Harman and Over, 2004). The task chosen for such evaluations must support a very high degree of interannotator agreement, i.e., consistent relevance decisions across subjects with respect to a predefined *gold standard*.

Unfortunately, a consistent gold standard has not yet been reported. For example, in two previous studies (Mani, 2001; Tombros and Sanderson, 1998), users' judgments were compared to "gold standard judgments" produced by members of the University of Pennsylvania's Linguistic Data Consortium. Although these judgments were supposed to represent the *correct* relevance judgments for each of the documents associated with an event, both studies reported that annotators' judgments varied greatly and that this was a significant issue for the evaluations. In the SUMMAC experiments, the Kappa score (Carletta, 1996; Eugenio and Glass, 2004) for interannotator agreement was reported to be 0.38 (Mani et al., 2002). In fact, large variations have been found in the initial summary scoring of an individual participant and a subsequent scoring that occurs a few weeks later (Mani, 2001; van Halteren and Teufel, 2003).

This paper attempts to overcome the problem of interannotator inconsistency by measuring summary effectiveness in an extrinsic task using a much more consistent form of user judgment instead of a gold standard. Using Relevance-Prediction increases the confidence in our results and strengthens the statistical statements we can make about the benefits of summarization.

The next section describes an alternative approach to measuring task-based usefulness, where the usage of external judgments as a gold standard is replaced by the user's own decisions on the full text. Following the lead of earlier evaluations (Oka and Ueda, 2000; Mani et al., 2002; Sakai and Sparck-Jones, 2001), we focus on relevance assessment as our extrinsic task.

[3] A topic is an event or activity, along with all directly related events and activities. An event is something that happens at some specific time and place, and the unavoidable consequences.

3 Evaluation of Usefulness of Summaries

We define a new extrinsic measure of task-based usefulness called *Relevance-Prediction*, where we compare a summary-based decision to the subject's own full-text decision rather than to a different subject's decision. Our findings differ from that of the SUMMAC results (Mani et al., 2002) in that using Relevance-Prediction as an alternative to comparision to a gold standard is a more realistic agreement measure for assessing usefulness in a relevance assessment task. For example, users performing browsing tasks must examine document surrogates, but open the full-text only if they expect the document to be interesting to them. They are not trying to decide if the document will be interesting to someone else.

To determine the usefulness of summarization, we focus on two questions:

- Can users make judgments on summaries that are consistent with their full-text judgments?
- Can users make judgments on summaries more quickly than on full document text?

First we describe the Relevance-Prediction measure for determining whether users can make accurate judgments with a summary. Following this, we describe our experiments and results using this measure, including the timing results of summaries compared to full documents.

3.1 Relevance-Prediction Measure

To answer the first question above, we define a measure called *Relevance-Prediction*, where subjects build their own "gold standard" based on the full-text documents. Agreement is measured by comparing subjects' surrogate-based judgments against their own judgments on the corresponding texts. The subject's judgment is assigned a value of *1* if his/her surrogate judgment is the same as the corresponding full-text judgment, and *0* otherwise. These values were summed over all judgments for a surrogate type and were divided by the total number of judgments for that surrogate type to determine the effectiveness of the associated summary method.

Formally, given a summary/document pair (s, d), if subjects make the same judgment on s that they did on d, we say $j(s, d) = 1$. If subjects change their judgment between s and d, we say $j(s, d) = 0$. Given a set of summary/document pairs DS_i associated with event i, the Relevance-Prediction score is computed as follows:

$$Relevance\text{-}Prediction(i) = \frac{\sum_{s,d \in DS_i} j(s, d)}{|DS_i|}$$

This approach provides a more reliable comparison mechanism than gold standard judgments provided by

other individuals. Specifically, Relevance-Prediction is more helpful in illuminating the usefulness of summaries for a real-world scenario, e.g., a browsing environment, where credit is given when an individual subject would choose (or reject) a document under both conditions. To our knowledge, this subject-driven approach to testing usefulness has never before been used.

3.2 Experiment Design

Ten human subjects were recruited to evaluate full-text documents and two summary types.[4] The original text documents were taken from the Topic Detection and Tracking 3 (TDT-3) corpus (Allan et al., 1999) which contains news stories and headlines, topic and event descriptions, and a mapping between news stories and their related topic and/or events. Although the TDT-3 collection contains transcribed speech documents, our investigation was restricted to documents that were originally text, i.e., newspaper or newswire, not broadcast news.

For our experiment we selected three distinct events and related document sets[5] from TDT-3. For each event, the subjects were given a description of the event (written by LDC) and then asked to judge relevance of a set of 20 documents associated with that event (using three different presentation types to be discussed below).

The events used from the TDT data set were events from world news occurring in 1998. It is possible that the subjects had some prior knowledge about the events, yet we believe that this would not affect their ability to complete the task. Subjects' background knowledge of an event can also make this task more similar to real-world browsing tasks, in which subjects are often familiar with the event or topic they are searching for.

The 20 documents were retrieved by a search engine. We used a constrained subset where exactly half (10) were judged relevant by the LDC annotators. Because all 20 documents were somewhat similar to the event, this approach ensured that our task would be more difficult than it would be if we had chosen documents from completely unrelated events (where the choice of relevance would be obvious even from a poorly written summary).

Each document was pre-annotated with the headline associated with the original newswire source. These headlines were used as the first summary type. We refer to them as HEAD (*Headline Surrogate*). The average length of the HEAD surrogates was 53 characters. In addition, we commissioned human-generated summaries[6] of each document as the second summary type; we refer to this as HUM (*Human Surrogate*). The average length of the HUM surrogates was 72 characters. Although neither of these summaries was produced automatically, our experiment allowed us to focus on the question of summary usefulness and to learn about the differences in presentation style as a first step toward experimentation with the output of automatic summarization systems.

Two main factors were measured: (1) differences in judgments for the three presentation types (HEAD, HUM, and the full-text document) and (2) judgment time. Each subject made a total of 60 judgments for each presentation type since there were 3 distinct events and 20 documents per event. To facilitate the analysis of the data, the subjects' judgments were constrained to two possibilities, *relevant* or *not relevant*.[7]

Although the HEAD and HUM surrogates were both produced by humans, they differed in style. The HEAD surrogates were shorter than the HUM surrogates by 26%. Many of these were "eye-catchers" designed to entice the reader to examine the entire document (i.e., purchase the newspaper); that is, the HEAD surrogates were not intended to stand in the place of the full document. By contrast, the writers of the HUM surrogates were instructed to write text that conveyed what happened in the full document. We observed that the HUM surrogates used more words and phrases extracted from the full documents than the HEAD surrogates.

Experiments were conducted using a web browser (Internet Explorer) on a PC in the presence of the experimenter. Subjects were given written and verbal instructions for completing their task and were asked to make relevance judgments on a practice event set. The judgments from the practice event set were not included in our experimental results or used in our analyses. The written instructions were given to aid subjects in determining requirements for relevance. For example, in an Election event documents describing new people in office, new public officials, change in governments or parliaments were suggested as evidence for relevance.

Each of the ten subjects made judgments on 20 documents for each of three different events. After reading each document or summary, the subjects clicked on a radio button corresponding to their judgment and clicked a *submit* button to move to the next document description. Subjects were not allowed to move to the next summary/document until a valid selection was made. No backing up was allowed. Judgment time was computed as the number of seconds it took the subject to read the full text document or surrogate, comprehend it, compare it to the event description, and make a judgment (timed up until the subject clicked the *submit* button).

[4] We required all human subjects to be native-English speakers to ensure that the accuracy of judgments was not degraded by language barriers.

[5] The three event and related document sets contained enough data points to achieve statistically significant results.

[6] The human summarizers were instructed to create a summary no greater than 75 characters for each specified full text document. The summaries were not compared for writing style or quality.

[7] If we allowed subjects to make additional judgments such as *somewhat relevant*, this could possibly encourage subjects to always choose this when they were the least bit unsure. Previous experiments indicate that this additional selection method may increase the level of variability in judgments (Zajic et al., 2004).

3.3 Order of Document/Surrogate Presentation

One concern with our evaluation methodology was the issue of possible memory effects or priming: if the same subjects saw a summary and a full document about the same event, their answers might be tainted. Thus, prior to the full experiment, we conducted pre-experiments (using 4 participants) with an extreme form of influence: we presented the summary and full text in immediate succession. In these experiments, we compared two document presentation approaches, termed "Drill Down" and "Complete Set." In the "Drill Down" document presentation approach all three presentation types were shown for each document, in sequence: first a single HEAD surrogate, followed by the corresponding HUM surrogate, followed by the full text document. This process was repeated 10 times.

In the "Complete Set" document-presentation approach we presented the complete set of documents using one surrogate type, followed by the complete set using another surrogate type, and so on. That is, the 10 HEAD surrogates were displayed all at once, followed by the corresponding 10 HUM surrogates, followed by the corresponding 10 full-text documents.

The results indicated that there was almost no effect between the two document-presentation approaches. The performance varied only slightly and neither approach consistently allowed subjects to perform better than the other. Therefore, we determined that the subjects were not associating a given summary with its corresponding full-text documents. This may be due, in part, to the fact that all 20 documents were related to the event—and according to the LDC relevance judgments half of these were actually about the same event.

Given that the variations were insignificant in these pre-experiments, we selected only the Complete-Set approach (no Drill-Down) for the full experiment. However, we still needed to vary the ordering for the two surrogate presentation types associated with each full-text document. Thus, each 20-document set was divided in half for each subject. In the first half, the subject saw the first 10 documents as: (1) HEAD surrogates, then HUM surrogates and then the full-text document; or (2) HUM surrogates, then HEAD surrogates, and then the full-text document. In the second half, the subject saw the alternative ordering, e.g., if a subject saw HEAD surrogates before HUM surrogates in the first half, he/she saw the HUM surrogates before HEAD surrogates for the second half. Either way, the full-text document was always shown last so as not to introduce judgment effects associated with reading the entire document before either surrogate type.

In addition to varying the ordering for the surrogate type, the ordering of the surrogates and full documents within the events were also varied. The subjects were grouped in pairs, and each pair viewed the surrogates and documents in a different order than the other pairs.

3.4 Experimental Hypotheses

We hypothesized that the summaries would allow subjects to achieve a Relevance-Prediction rate of 70–90%. Since these summaries were significantly shorter than the original document text, we expected that the rate would not be 100% compared to the judgments made on the full document text. However, we expected higher than a 50% ratio, i.e., higher than that of random judgments on all of the surrogates. We also expected high performance because the meaning of the original document text is best preserved when written by a human (Mani, 2001).

A second hypothesis is that the HEAD surrogates would yield a significantly lower agreement rate than that of the HUM surrogates. Our commissioned HUM surrogates were written to stand in place of the full document, whereas the HEAD surrogates were written to catch a reader's interest. This suggests that the HEAD surrogates might not provide as informative a description of the original documents as the HUM surrogates.

We also tested a third hypothesis: that our Relevance-Prediction measure would be more reliable than that of the *LDC-Agreement* method used for SUMMAC-style evaluations (thus providing a more stable framework for evaluating summarization techniques). LDC-Agreement compares a subject's judgment on a surrogate or full text against the "correct" judgments as assigned by the TDT corpus annotators (Linguistic Data Consortium 2001).

Finally, we tested the hypothesis that using a text summary for judging relevance would take considerably less time than using the corresponding full-text document.

4 Experimental Results

Table 1 shows the subjects' judgments using both Relevance-Prediction and LDC-Agreement for each of three events. Using our Relevance-Prediction measure, the HUM surrogates yield averages between 79% and 86%, with an overall average of 81%, thus confirming our first hypothesis.

However, we failed to confirm our second hypothesis. The HEAD Relevance-Prediction rates were between 71% and 82%, with an overall average of 76%, which was lower than the rates for HUM, but the difference was not statistically significant. It appeared that subjects were able to make consistent relevance decisions from the non-extractive HEAD surrogates, even though these were shorter and less informative than the HUM surrogates.

A closer look reveals that the HEAD summaries sometimes contained enough information to judge relevance, yielding almost the same number of true positives (and true negatives) as the HUM summaries. For example, a document about the formation of a coalition government to avoid violence in Cambodia has the HEAD surrogate *Cambodians hope new government can avoid past mistakes*. By contrast, the HUM surrogate for this same event was *Rival parties to form a coalition government to avoid violence in Cambodia*. Although the HEAD surrogate

Surrogate	EVENT 1		EVENT 2		EVENT 3		Overall Avg		Avg Time
	LDC	RP	LDC	RP	LDC	RP	LDC	RP	(seconds)
HEAD	67%	76%	66%	71%	70%	82%	67%	76%	4.60
HUM	69%	80%	73%	86%	62%	79%	68%	81%	4.57
DOC	—	—	—	—	—	—	—	—	13.38

Table 1: Relevance-Prediction (RP) and LDC-Agreement (LDC) Rates for HEAD and HUM Surrogates for each Event

uses words that do not appear in the original document (*hope* and *mistakes*), the subject may infer the relevance of this surrogate by relating *hope* to the notion of forming a coalition government and *mistakes* to violence.

On the other hand, we found that the lower degree of informativeness of HEAD surrogates gave rise to over 50% more false negatives than the HUM summaries. This statistically significant difference will be discussed further in Section 6.

As for our third hypothesis, Table 1 illustrates a substantial difference between the two agreement measures. For each of the three events, the Relevance-Prediction rate is at least five percent higher than that of the LDC-Agreement approach, with an average of 8.8% increase for the HEAD summary and a 13.3% average increase for the HUM summary. The average rates across events show a statistically significant difference between LDC-Agreement and Relevance-Prediction for both HUM summaries with $p<0.01$ and HEAD summaries with $p<0.05$. This significance was determined through use of a single factor ANOVA statistical analysis. The higher Relevance-Prediction rate supports our statement that this approach provides a more stable framework for evaluating different summarization techniques.

Finally, the average timing results shown in Table 1 confirm our fourth hypothesis. The subjects took 4-5 seconds (on average) to make judgments on both the HEAD and HUM summaries, as compared to about 13.4 seconds to make judgments on full text documents. This shows that it takes subjects almost 3 times longer to make judgments on full text documents as it took to make judgments on the summaries (HEAD and HUM). This finding is not surprising since text summaries are an order of magnitude shorter than full-text documents.

5 Correlation with Intrinsic Evaluation Metric: ROUGE

We now turn to the task of correlating our extrinsic task performance with scores produced by an intrinsic evaluation measure. We used the Recall Oriented Understudy for Gisting Evaluation (ROUGE) metric version 1.2.1. In previous studies (Dorr et al., 2004) ROUGE was shown to have a very low correlation with the LDC-Agreement measurement results of the extrinsic task. This was attributed to low interannotator agreement in the gold standard. Our goal was to test whether our new Relevance-Prediction technique would allow us to induce higher correlations with ROUGE.

5.1 Extrinsic Agreement Data

To reduce the effect of outliers on the correlation between ROUGE and the human judgments, we averaged over all judgments for each subject (20 judgments $\times$ 3 events) to produce 60 data points. These data points were then partitioned into either 1, 2, or 4 partitions of equal size. (Partitions of size four have 15 data points, partitions of size two have 30 data points, and partitions of size one have 60 data points per subject—or a total of 600 datapoints across all 10 subjects). To ensure that the correlation did not depend on a specific partition, we repeated this same process using 10,000 different (randomly generated) partitions for each of the three partition sizes.

Partitioned data points of size four provided a high degree of noise reduction without compromising the size of the data set (15 points). Larger partition sizes would result in too few data points and compromise the statistical significance of our correlation results. In order to show the variation within a single partition, we used the partitioning of size 4 with the smallest mean square error on the human headline compared to the other partitionings as a representative partition. For this representative partitioning, the individual data points P1–P15 of that partition are shown for each of the two agreement measures in Tables 2 and 3. This shows that, across partitions, the maximum and minimum Relevance-Prediction rates for HEAD (93% and 60%) are higher than the corresponding LDC-Agreement rates (85% and 50%). The same trend is seen with the HUM surrogates: Relevance-Prediction maximum of 98%, minimum of 68%; and LDC-Agreement maximum 88%, minimum of 55%.

5.2 Intrinsic ROUGE Score

To correlate the partitioned agreement scores above with our intrinsic measure, we first ran ROUGE on all 120 surrogates in our experiment (i.e., the HUM and HEAD surrogates for each of the 60 event/document pairs) and then averaged the ROUGE scores for all surrogates belonging to the same partitions (for each of the three partition sizes). These partitioned ROUGE values were then used for detecting correlations with the corresponding partitioned agreement scores described above.

Table 4 shows the ROUGE scores, based on 3 reference summaries per document, for partitions P1–P15 used in the previous tables.[8] For brevity, we include

[8]We commissioned a total of 180 human-generated reference summaries (3 for each of 60 documents) (in addition to the human generated summaries used in the experiment).

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
HEAD	80%	80%	85%	70%	73%	60%	80%	75%	60%	75%	88%	68%	80%	93%	83%
HUM	83%	88%	85%	68%	75%	75%	93%	75%	98%	90%	75%	70%	80%	90%	78%

Table 2: Relevance-Prediction Rates for HEAD and HUM Surrogates (Representative Partition of Size 4)

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
HEAD	70%	73%	85%	70%	63%	60%	60%	85%	50%	73%	70%	78%	65%	63%	73%
HUM	68%	75%	58%	68%	75%	70%	68%	80%	88%	58%	63%	55%	55%	60%	78%

Table 3: LDC-Agreement Rates for HEAD and HUM Surrogates (Representative Partition of Size 4)

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Avg
HEAD	.10	.23	.13	.27	.20	.24	.26	.22	.13	.08	.30	.16	.26	.27	.30	.211
HUM	.16	.22	.17	.23	.19	.36	.39	.29	.28	.25	.37	.22	.22	.39	.27	.269

Table 4: Average Rouge-1 Scores for HEAD and HUM Surrogates (Representative Partition of Size 4)

only ROUGE 1-gram measurement (R1).[9] The ROUGE scores for HEAD surrogates were slightly lower than those for HUM surrogates. This is consistent with our statements earlier about the difference between non-extractive "eye-catchers" and informative headlines. Because ROUGE measures whether a particular summary has the same words (or n-grams) as a reference summary, a more constrained choice of words (as found in the extractive HUM surrogates) makes it more likely that the summary would match the reference.

A summary in which the word choice is less constrained—as in the non-extractive HEAD surrogates—is less likely to share n-grams with the reference. Thus, we may see non-extractive summaries that have almost identical meanings, but very different words. This raises the concern that ROUGE may be sensitive to the style of summarization that is used. Section 6 discusses this point further.

5.3 Intrinsic and Extrinsic Correlation

To test whether ROUGE correlates more highly with Relevance-Prediction than with LDC-Agreement, we calculated the correlation for the results of both techniques using Pearson's *r* (Siegel and Castellan, 1988):

$$\frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

where r_i is the ROUGE score of surrogate i, $\bar{r}$ is the average ROUGE score of all data points, s_i is the agreement score of summary i (using Relevance-Prediction or LDC-Agreement), and $\bar{s}$ is the average agreement score. Pearson's statistics is commonly used in summarization and machine translation evaluation, see e.g. (Lin, 2004; Lin and Och, 2004).

As one might expect, there is some variability in the correlation between ROUGE and human judgments for the different partitions. However, the boxplots for both HEAD and HUM indicate that the first and third quartile were relatively close to the median (see Figure 1).

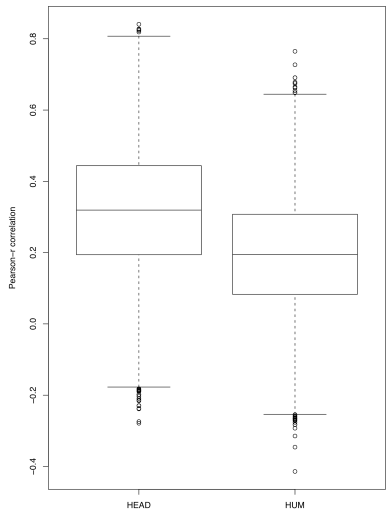


Figure 1: Distribution of the Correlation Variation for Relevance-Prediction on HEAD and HUM

Table 5 shows the Pearson Correlations with ROUGE-1 using Relevance-Prediction and LDC-Agreement. For Relevance-Prediction, we observed a positive correlation for both surrogate types, with a slightly higher correlation for HEAD than HUM. For LDC-Agreement, we observed no correlation (or a minimally negative one) with ROUGE-1 scores, for both the HEAD and HUM surrogates. The highest correlation was observed for Relevance-Prediction on HEAD.

We conclude that ROUGE correlates more highly with the Relevance-Prediction measurement than the LDC-Agreement measurement, although we should add that none of the correlations in Table 5 were statistically significant at $p < 0.05$. The low LDC-Agreement scores are consistent with previous studies where poor correlations

[9] We also computed ROUGE 2-gram, ROUGE L and ROUGE W, but the trend for these did not differ from ROUGE-1.

Surrogate	P = 1	P = 2	P = 4
HEAD (RP)	0.1270	0.1943	0.3140
HUM (RP)	0.0632	0.1096	0.1391
HEAD (LDC)	-0.0968	-0.0660	-0.0099
HUM (LDC)	-0.0395	-0.0236	-0.0187

Table 5: Pearson Correlations with ROUGE-1 for Relevance-Prediction (RP) and LDC-Agreement (LDC), where Partition size (P) = 1, 2, and 4

were attributed to low interannotator agreement rates.

6 Discussion

Our results suggest that ROUGE may be sensitive to the style of summarization that is used. As we observed above, many of the HEAD surrogates were not actually summaries of the full text, but were eye-catchers. Often, these surrogates did not allow the subject to judge relevance correctly, resulting in lower agreement. In addition, these same surrogates often did not use a high percentage of words that were actually from the story, resulting in low ROUGE scores. (We noticed that most words in the HUM surrogates appeared in the corresponding stories.) There were three consequences of this difference between HEAD and HUM: (1) The rate of agreement was lower for HEAD than for HUM; (2) The average ROUGE score was lower for HEAD than for HUM; and (3) The correlation of ROUGE scores with agreement was higher for HEAD than for HUM.

A further analysis supports the (somewhat counterintuitive) third point above. Although the ROUGE scores of true positives (and true negatives) were significantly lower for HEAD surrogates (0.2127 and 0.2162) than for HUM surrogates (0.2696 and 0.2715), the number of false negatives was substantially higher for HEAD surrogates than for HUM surrogates. These cases corresponded to much lower ROUGE scores for HEAD surrogates (0.1996) than for HUM (0.2586) surrogates.

A summary of this analysis is given in Table 6, where true positives and negatives are indicated by Rel/Rel and NonRel/NonRel, respectively, and false positives and negatives are indicated by Rel/NonRel and NonRel/Rel, respectively.[10] The numbers in parentheses after each ROUGE score refer to the standard deviation for that score. This was computed as follows:

$$Std.\text{-}Dev. = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

where N is the number of surrogates in a particular judgment category (e.g., $N = 245$ for the HEAD-based NonRel/Rel judgments), x_i is the ROUGE score for the i^{th} surrogate, and $\bar{r}$ is the average of all ROUGE scores in that category.

Although there were very few false positives (less than 6% for both HEAD and HUM), the number of false negatives (NonRel/Rel) was particularly high for HEAD (50% higher than for HUM). This difference was statistically significant at $p<0.01$ using the t-test. The large number of false negatives with HEAD may be attributed to the eye-catching nature of these surrogates. A subject may be misled into thinking that this surrogate is not related to an event because the surrogate does not contain words from the event description and is too broad for the subject to extract definitive information (e.g., the surrogate *There he goes again!*). Because the false negatives were associated with the lowest average ROUGE score (0.1996), we speculate that, if a correlation exists between Relevance-Prediction and ROUGE, the false negatives may be a major contributing factor.

Based on this experiment, we conjecture that ROUGE may not be a good method for measuring the usefulness of summaries when the summaries are not extractive. That is, if someone intentionally writes summaries that contain different words than the story, the summaries will also likely contain different words than a reference summary, resulting in low ROUGE scores. However, the summaries, if well-written, could still result in high agreement with the judgments made on the full text.

7 Conclusion

We have shown that two types of human summaries, HEAD and HUM, can be useful for relevance assessment in that they help a user achieve 70-85% agreement in relevance judgments. We observed a 65% reduction in judgment time between full texts and summaries. These findings are important in that they establish the usefulness of summarization and they support research and development of additional summarization methods, including automatic methods.

We introduced a new method for measuring agreement, *Relevance-Prediction*, which takes a subject's full-text judgment as the standard against which the same subject's summary judgment is measured. Because Relevance-Prediction was more reliable than LDC-Agreement judgments, we encourage others to use this measure in future summarization evaluations.

Using this new method, we were able to find positive correlations between relevance assessments and ROUGE scores for HUM and HEAD surrogates, where only

[10] We also included (average) elapsed times for summary judgments in each of the four categories. One might expect a "relevant" judgment to be much quicker than a "non-relevant" judgment (since the latter might require reading the full summary). However, it turned out non-relevant judgments did not always take longer. In fact, the NonRel/NonRel cases took considerably less time than the Rel/Rel and Rel/NonRel cases. On the other hand, the NonRel/Rel cases took considerably more time—almost as much time as reading the full text documents—an indication that the subjects may have re-read the summary a number of times, perhaps vacillating back and forth. Still, the overall time savings was significant, given that the vast majority of the non-relevant judgments were in the NonRel/NonRel category.

Judgment (Surr/Doc)	HEAD			HUM		
	Raw	R1-Avg	Avg Time	Raw	R1-Avg	Avg Time
Rel/Rel	211 (35%)	0.2127 (±0.120)	4.6	251 (42%)	0.2696 (±0.130)	4.2
Rel/NonRel	27 (5%)	0.2115 (±0.110)	7.1	35 (6%)	0.2725 (±0.131)	4.6
NonRel/Rel	117 (19%)	0.1996 (±0.127)	8.5	77 (13%)	0.2586 (±0.120)	13.8
NonRel/NonRel	245 (41%)	0.2162 (±0.126)	2.5	237 (39%)	0.2715 (±0.131)	1.9
TOTAL	600 (100%)	0.2115 (±0.124)	4.6	600 (100%)	0.2691 (±0.129)	4.6

Table 6: Subjects' Judgments and Corresponding Average ROUGE 1 Scores

negative correlations were found using LDC-Agreement scores. We found that both the Relevance-Prediction and the ROUGE-1 scores were higher for human-generated summaries than for the original headlines. It appears that most of the difference is induced by surrogates that are eye-catchers (rather than true summaries), where both agreement and ROUGE scores are low.

Our future work will include further experimentation with automatic summarization methods to determine the level of Relevance-Prediction. We aim to determine how well automatic summarizers help users complete tasks, and to investigate which automatic summarizers perform better than others. We also plan to test for correlations between ROUGE and human task performance with automatic summaries, to further investigate whether ROUGE is a good predictor of human task performance.

Acknowledgements

This work was supported in part by DARPA TIDES Cooperative Agreement N66001-00-2-8910.

References

James Allan, Hubert Jin, Martin Rajman, Charles Wayne, Daniel Gildea, Victor Lavrenko, Rose Hoberman, and David Caputo. 1999. Topic-based Novelty Detection. Technical Report 1999 Summer Workshop at CLSP Final Report, Johns Hopkins, Maryland.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Lingusitics*, 22(2):249–254, June.

Bonnie J. Dorr, Christof Monz, Douglas Oard, Stacy President, and David Zajic. 2004. Extrinsic Evaluation of Automatic Metrics for Summarization. Technical report, University of Maryland, College Park, MD. LAMP-TR-115, CAR-TR-999, CS-TR-4610, UMIACS-TR-2004-48.

Barbara Di Eugenio and Michael Glass. 2004. Squibs and Discussions - The Kappa Statistic: A Second Look. *Computational Linguistics*, pages 95–101.

Donna Harman and Paul Over. 2004. *Proceedings of the DUC 2004*. Boston, MA.

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI Symposium on Intelligent Summarization*, Stanford University, CA, March 23-25.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of HLT-NAACL 2003 Workshop*, pages 71–78, Edmonton Canada, May-June.

Chin-Yew Lin and Franz Joseph Och. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, August 23–27.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25–26.

I. Mani and E. Bloedorn. 1998. Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, 1(1):35–67.

Inderjeet Mani, Gary Klein, David House, and Lynette Hirschman. 2002. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Inderjeet Mani. 2001. Summarization Evaluation: An Overview. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the NAACL 2004*, Boston, MA.

Mamiko Oka and Yoshihiro Ueda. 2000. Evaluation of Phrase-Representation Summarization Based on an Information Retrieval Task. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 59–68, New Brunswick, NJ.

Dragomir Radev and Kathleen McKeown. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, pages 469–500.

Tetsuya Sakai and Karen Sparck-Jones. 2001. Generic Summaries for Indexing in Information Retrieval - Detailed Test Results. Technical Report TR513, Computer Laboratory, University of Cambridge.

Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition.

Karen Sparck-Jones and J.R. Gallier. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, Berlin.

Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10.

Hans van Halteren and Simone Teufel. 2003. Examining the Consensus Between Human Summaries: Initial Experiments with Factoid Analysis. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*.

David Zajic, Bonnie J. Dorr, Richard Schwartz, and Stacy President. 2004. Headline Evaluation Experiment Results. Technical report, University of Maryland, College Park, MD. UMIACS-TR-2004-18.

On the Subjectivity of Human Authored Short Summaries

BalaKrishna Kolluru **Yoshihiko Gotoh**
University of Sheffield, Department of Computer Science
Sheffield S1 4DP, United Kingdom
{b.kolluru, y.gotoh}@dcs.shef.ac.uk

Abstract

We address the issue of human subjectivity when authoring summaries, aiming at a simple, robust evaluation of machine generated summaries. Applying a cross comprehension test on human authored short summaries from broadcast news, the level of subjectivity is gauged among four authors. The instruction set is simple, thus there is enough room for subjectivity. However the approach is robust because the test does not use the absolute score, relying instead on relative comparison, effectively alleviating the subjectivity. Finally we illustrate the application of the above scheme when evaluating the informativeness of machine generated summaries.

1 Introduction

Subjectivity plays an important role when removing the unwanted or redundant information for summarising a document. Human beings tend to disagree on what should be a 'one good summary' (Mani, 2001). This is probably because every individual, whilst arriving at a summary, looks at things from a different perspective. Guided by various factors such as educational background, profession, personal interests and experience, an individual decides whether a certain aspect is worth being included in a summary. What might seem relevant to one person could be deemed redundant by another when reading the same story, thus accounting for more than one 'correct' summary. The issue of subjectivity gains prominence as the compression ratio increases, *i.e.*, the shorter the summary, the larger the number of 'correct' summaries (Lin and Hovy, 2003b). This is due to the fact that assimilation of seemingly important contents takes priority while discarding the redundant information. This is a highly subjective aspect.

Although the subjectivity reflects individual's thoughts, there will also be some information commonly observed in different summaries of the same story. Stated otherwise, words in a summary may vary, phrases may vary, and often the grammatical structure may not be the same, but a certain degree of information may be common across summaries. To what degree is information uniform across different summaries? How much subjectivity is there? How do we account for similar information stated using different words, expressions, or grammatical structure when comparing summaries? How does this help when gauging the informativeness? Does the subjectivity cause any adverse effects when evaluating summaries? It is these questions that we aim to address in this paper.

Let us assume that the atomic facts of a summary account for its relevance. Then, a simple question that elicits any one of these atomic facts represents a benchmark for assessing its informativeness. We wish to evaluate the quality of a summary in terms of atomic facts commonly observed in-, or subjectively discarded from, assorted human authored short summaries. In our quest to quantify the subjectivity, we devise a cross comprehension test along the lines of (Hirschmann et al., 1999) for extracting atomic contents. The comprehension test is modelled on a question-answer style framework. 'Crossing' the model turns out to be an effective scheme for measuring the divergence among multiple summaries. Questions are prepared by the subject who wrote the original summary (Section 3). Their answers

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 9–16, Ann Arbor, June 2005.

should be derived by reading the summary alone. Summary-questionnaire pairs are then swapped in such a way that any summary is paired with questions written by other subjects (Section 4). The number of questions that cannot be answered by reading the summary accounts for the subjectiveness of the author (Section 5). Finally, we address how the cross comprehension test can be used for evaluating machine generated summaries (Section 6).

2 Related Works

There have been a number of studies concerned with collating and analysing of human authored summaries, with the aim of producing and evaluating machine generated summaries. A phrase weighting process called the 'pyramid method' was described in (Nenkova and Passonneau, 2004). They exploited the frequency of the same (similar) information that was in multiple summaries of the same story. It was referred to as a *summarisation content unit* (SCU). Increasing stability of pyramid scores was observed as the pyramid grew larger. The authors concluded, however, that the initial creation of the pyramid was a tedious task because a large number of SCUs had to be hand annotated.

In (Van Halteren and Teufel, 2003), the co-occurrence of atomic information elements, called *factoids*, was examined whilst analysing 50 different summaries of two stories. A candidate summary was compared with the reference using factoids in order to measure the informativeness. The authors observed that from a wide selection of factoids only a small number were included in all summaries. From a pool of factoids, approximately 30% were taken to build a consensus summary that could be used as a 'gold standard'.

Summary evaluation has been recognised as a sensitive, non-trivial task. In (Radev and Tam, 2003) the *relative utility* was calculated based on a significance ranking assigned to each sentence. A word network based summary evaluation scheme was proposed in (Hori et al., 2003), where the accuracy was weighted by the posterior probability of the manual summaries in the network. Significantly, they surmised the independence of their criterion from the variations in hand summaries.

A regression analysis was performed in (Hirohata et al., 2005) and concluded that objective evaluations were more effective than subjective approaches. Although their experiments were concerned with presentation speech, the results do have a universal appeal.

Another notable development in the field is the n-gram co-occurrence matching technique as proposed in (Lin and Hovy, 2003a). Their tool, ROUGE, compares the number of n-gram matches between a reference and a machine generated summary. Recently, ROUGE was piloted for evaluation of summaries from newspaper/newswire articles (Over and Yen, 2004). ROUGE simulated the manual evaluation well for that task, although it is still unclear how closely it well to other tasks.

To some extent, the work described in this paper is close to that of (Nenkova and Passonneau, 2004) and (Van Halteren and Teufel, 2003). We analyse human authored summaries associating human subjectivity with their unique interpretation of stories. We consider their effect when evaluating machine generated summaries.

3 Production of Human Authored Short Summaries

Our aim is to investigate an effective, robust approach to summary evaluation. In this paper, we identify and quantify the aspect of human subjectivity while authoring short summaries. To this end, four subjects produced a short summary (approximately 100 characters, or 15 words) for broadcast news stories given a simple instruction set. This summary is referred to as a 'one line' summary because it corresponds approximately to the average sentence length for this data set.

3.1 Author Profiles

Four summary authors are briefly profiled below:

Subject A. A linguist by profession, a polyglot out of interest, and an author by hobby. This subject is fluent in English, Spanish and French; English being the first language. The subject is trained to write summaries and translations.

Subject B. A manager by qualification and a polyglot by necessity; English is a second language. This subject was trained in making presentations and documentation. We hoped to benefit from the synergy

of both fields for summary production.

Subject C. A physicist by qualification and currently working towards a PhD in speech recognition. English is the first language. In addition, this subject has an interest in theatre and drama, thus is exposed to literature and related fields.

Subject D. Working on research in multiparty meetings as a post doctoral fellow. English is the first language for this subject. Experience of meeting summarisation.

All subjects are educated to at least graduate level, and have are fluent in English. It was expected that they could produce summaries of good quality without detailed instruction or further training. A simple instruction set (discussed later) was given, leaving wide room for interpretation about what might be included in the summary. Hence subjectivity was promoted.

3.2 Data

The human subjects worked on a small subset of American broadcast news stories from the TDT-2 corpus (Cieri et al., 1999). They were used for NIST TDT evaluations and the TREC-8 and TREC-9 spoken document retrieval evaluations. Each program in the corpus contained 7 to 8 news stories on average, spanning 30 minutes as broadcast which might be reduced to 22 minutes once advertisement breaks were removed. A set of 51 hand transcriptions were manually selected from the corpus. The average length was 487 words in 25 sentences per transcription.

3.3 Instructions

Summary production. A simple instruction was given to the subjects in order to arrive at a summary:

- Each summary should contain about 100 characters, possibly in the subject's own words.

As the news stories ranged from 16 to 84 sentences, subjects would have to prioritise information that could be included in their 'one line' summary. The instruction implicitly encouraged the subjects to put as much important information as possible into a summary, while maintaining a good level of fluency. It was also a flexible instruction so that subjects were able to use their own expressions when necessary. After completion of the task, they commented that this instruction made them experiment with different words to shorten or expand the information they wanted to include. For example, how could an earthquake disaster be expressed in different ways:

> 8000+ feared dead? . . . or
> thousands of people killed? . . . or
> a lot of people are believed to be dead?

Another feature of this instruction was the amount of generalisation that a subject was likely to use. For example, a subject could say

> US Senate to decide on tobacco bill

but given the length constraints, it could be like

> Senate to vote on bill, hiking tobacco price

while adding extra information, but omitting specific details.

Questionnaire production. When producing summaries, subjects were aware that they also had to prepare questions with the following instructions:

- A questionnaire may consists of 2–4 questions;
- An answer must be found in the particular summary, without reading the entire story;
- Yes / no questions should not be used;
- The summary may roughly be reconstructed from the question-answer set.

Each fact might be questioned in such a way that the particular summary could be recovered. Ideally we would expect each question to elicit a precise information point chosen for the summary — *e.g.*, who did it, when did it happen, what was the cause? The question-answer set enabled us to gauge the most relevant information as decided by the subjects, so that their subjectiveness became apparent.

3.4 Full Sample

A 'one line' summary-questionnaire pair was produced for 51 broadcast news stories by each of the four subjects. The statistics in Table 1 show the average number of words and characters for each summary. It is observed that Subjects **A** (6.1 characters / word) and **C** (5.8) tended to use longer words than **B**

Subject	#words	#characters	#questions
A	16	113	3.7
B	17	99	3.5
C	12	81	2.4
D	21	131	3.0

Table 1: This table shows the average number of words and characters for each summary, and the average number of questions per summary.

(4.9) and **D** (5.3). The table also shows how the average number of questions varies between subjects.

Table 2 shows a full sample. The complete news story is found in the Appendix. The difference between the four summaries can be clearly observed. One noticeable aspect is the amount of abstraction preferred by various subjects. Both Subjects **A** and **D** fully utilised words from the news story and made a small amount of abstraction. In particular, Subject **A** chose to pick out a person ('*Fisher*') who conducted the study, while **D** opted for specifics of the study ('*dopamine*' — a responsible chemical). On the other hand, Subjects **B** and **C** have rendered their interpretation of the story in their own expressions. They have produced a highly abstracted summary reflecting the sense of the story while ignoring the specifics — nevertheless they were very different from each other. All four summaries happen to be of good quality, however it is the sheer divergence in the words, the expressions and subjective interpretation that is striking.

Word usage among the subjects is also interesting — *e.g.*, '*visual images*' as against '*physical traits*'; similarly '*inner feelings*' as against '*chemistry*'. Such expressions and idioms are open for interpretation, making it difficult to quantify the informativeness of any summary.

There also exist many factual news stories among the 51 test stories. It is left for a future study to compare between factual and non-factual news, in particular about the amount of abstraction.

4 Cross Comprehension Test

Each question can extract a relevant answer from the particular summary by the same author. If a question set were applied to a different summary, some answers may be discernible whereas others may not. The cross comprehension test achieves this by swapping a summary-questionnaire pair, *i.e.*, each summary was paired with questions produced by different authors. Figure 1 illustrates the way it works.

Subject A
Summary:
Fisher's study claims we seek partners using unconscious love maps; women prefer status, men go for physical traits.
Questions:
1. Who is the author of this study?
2. What claim does the researcher make concerning our method for seeking a sexual partner?
3. What do women look for in men?
4. What do men go for?

Subject B
Summary:
Internal feelings of love between men and women are unique; external features depend on culture.
Questions:
1. What are unique?
2. What is this topic about?
3. What differs between men and women?
4. Why does it differ?

Subject C
Summary:
Culture and chemistry both play a role in the science of romance.
Questions:
1. What is being discussed?
2. What are the factors affecting the particular event?

Subject D
Summary:
Men are turned on by visual images and women are more focused on someone's character traits, based on dopamine.
Questions:
1. What do women look for in men?
2. What do men look for in women?
3. What is the chemical that controls attraction?

Table 2: Summary-questionnaire pairs produced from broadcast news stories by four subjects.

A single judge examines whether each question can be answered by reading a swapped summary. The judge is a person different from the four summary authors. Further, if the answer is found, it may be relevant, partially relevant, or totally irrelevant to the one expected by the author. Thus, the decision is made from the following four options:

relevant: a relevant answer is found — the answer is deemed to be relevant if it conveys the same meaning as expected by the author even if a different expression is used;

partially relevant: an answer is partially relevant;

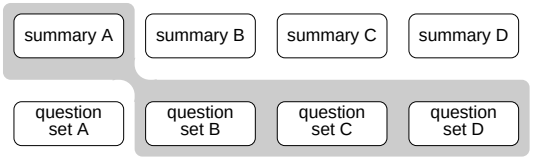


Figure 1: The cross comprehension test swaps summary-questionnaire pairs between subjects. For example, a summary by Subject **A** may be questioned by those set by Subjects **B**, **C**, and **D**.

irrelevant: an answer is found, but is totally different from that expected by the author.

not found: no answer is found.

Sample (re-visited). Table 3 shows the summary and questions crossed from the sample in Table 2. For example, when the 'one line' summary authored by Subject **A** is matched with Subject **B**'s questions, corresponding answers may be

1. ?;
2. seeking partners;
3. women prefer status, men go for physical traits;
4. unconscious love maps.

We may thus conclude answers are '*not found*', '*relevant*', '*irrelevant*', and '*partially relevant*' because, from Table 2, actual answers sought by **B** were

1. internal feelings;
2. love between men and women;
3. external features;
4. cultural reason.

Compensating ill-framed questions. We are aware that not all 'one line' summaries were well written. For example, it may be difficult to reach the expected answer ('external features') for Question 3 by Subject **B** ('What differs between men and women?') by reading the summary from the same subject. Moreover, subjects occasionally set a question that could not be answered properly by reading the particular summary alone. By crossing the summary-questionnaire pair, ill-framed questions are effectively compensated, because they are equally posed to all candidate summaries.

Judgement difficulty. One potential problem in this scheme is the difficulty a judge may face when choosing from the four options. A judge's decision can also be affected by subjectivity. Our assumptions are that (1) because there are only four options, there is less room for the subjectivity in comparison

Summary by Subject A:
Fisher's study claims we seek partners using unconscious love maps; women prefer status, men go for physical traits.
Questions by Subject B:
1. What are unique? (N)
2. What is this topic about? (R)
3. What differs between men and women? (I)
4. Why does it differ? (P)
Questions by Subject C:
1. What is being discussed? (R)
2. What are the factors affecting the particular event? (R)
Questions by subject D:
1. What do men look for in women? (R)
2. What do women look for in men? (R)
3. What is the chemical that controls attraction? (N)

Table 3: What if the summary by Subject **A** is questioned by Subjects **B**, **C**, or **D**? (R), (P), (I), and (N) after each question indicate the answer is *relevant*, *partially relevant*, *irrelevant*, and *not found*.

to the summary writing task, and that (2) a decision between '*relevant*' and '*partially relevant*' and one between '*irrelevant*' and '*not found*' are both not very important because the former two are roughly associated with commonly shared information and the latter two correspond to the subjective part. Although the following section shows results by a single judge, we are currently conducting the same experiments using multiple judges in order to quantify our assumptions.

5 Evaluation Results

Each of the four 'one line' summaries from the 51 broadcast news stories were evaluated using three sets of 'crossed' questions.

5.1 Summary Relevance

Figure 2(a) shows, when paired with questions by other subjects, how many answers could be found in a candidate summary. The figure indicates that summaries authored by the different subjects contained '*relevant*' information for less than half (47% overall average for four subjects) of questions. The number goes up slightly (61%) if '*partially relevant*' answers are included. The number of answers that were '*not found*' indicates the level of subjectivity for this 'summary writing' exercise; more than one third (35%) of information that one subject thought

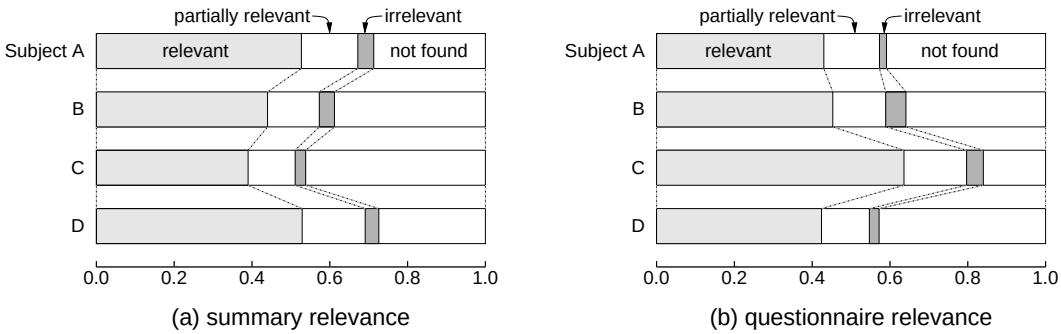


Figure 2: Summary relevance was measured when evaluated against questions by other subjects, while questionnaire relevance was calculated when evaluated against summaries by other subjects.

was the most important was discarded by the others. We surmise that '*irrelevant*' answers were also caused by the subjectivity; occasionally authors arrived at contradictory summaries of the same story due to its ambiguous nature. In such cases, questions were produced from that author's subjective view, and they certainly affected the relevance of a summary by the other subject.

Another notable outcome of this experiment is that the number of answers found '*relevant*', '*partially relevant*' or '*irrelevant*' was 71%, 61%, 54% and 73% for Subjects **A**, **B**, **C**, and **D**, respectively. This seems roughly proportional to the average length of summaries by each subject (113, 99, 81, and 131 characters, respectively). The longer the summary, the more information one can write in the summary. It is thus hypothesised that only the summary length matters for finding the '*relevant*' information in summaries. Looking at this outcome from a different perspective, there is no evidence that one author was more subjective than the others.

5.2 Questionnaire Relevance

Figure 2(b) shows, when paired with summaries by other subjects, how many candidate questions could be answered. It is based on the same evaluation as 2(a), but observed from the different angle. Approximately the same number (55–59%) of '*relevant*', and '*partially relevant*' answers were found for Subjects **A**, **B**, and **D**. However, it was much higher (80%) for Subject **C**. The reason seems to be that this subject frequently set questions that might accept a wide range of answers, while other subjects tended to frame questions that required more specific information in the summary; *e.g.*, Subject **C**'s '*what is being discussed?*' was a general question that was more likely to have some answer than Subject **B**'s question '*what differs between men and women?*'.

5.3 Discussion

The overall number of '*relevant*' and '*partially relevant*' answers found by the cross comprehension test was just over 61% for four subjects. This accounts for the amount of information that was agreed by all the subjects as important. For more than one third of summary contents, subjects had different opinions about whether they should be in their 'one line' summaries, resulting in categories such as '*irrelevant*' or '*not found*'. Occasionally these categories resulted from ill-framed questions, but such questions were infrequent. For most of the cases, they were caused by the subjectivity of a different individual.

We noted earlier that only the summary length matters and there is no evidence that one author was more subjective than the others. It is probably because, given a clear instruction about the summary length (*i.e.*, roughly 100 characters for this task), there is an upper bound for the amount of information that anyone can fit into the summary, while maintaining fluency. When the summary is short, one has to make a serious decision about which important information should go into a summary, and the decision often reflects one's subjective thoughts. Our argument is that, assuming the subject's effort, the amount of subjectivity was controlled by the summary length constraints rather than an individual's nature.

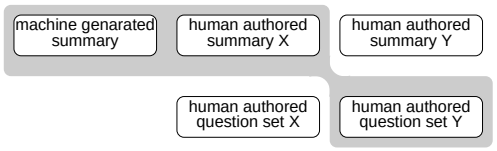


Figure 3: Evaluation of machine generated summaries by the cross comprehension test.

The diversity of summaries caused by individual subjectivity may be alleviated by carefully drafting an instruction set. However it probably results in a large list of instructions, and the drafting process certainly will not be straightforward. Further, it is not likely that we can ever completely remove the subjectivity from human work. Indeed, if subjectivity disappeared from human authored summary by well crafted instructions, it would be more like turning human activity into a mechanical process, rather than a machine to simulate human work.

A non-trivial problem of the approach may be the amount of human effort needed for evaluation. Production of summary-questionnaire pairs may not be difficult, as it is based on a simple instruction set and even accepts ill-framed questions, but it still requires human time. On the other hand, a judge's role is the most critical — it is labour intensive, and the effect of potentially subjective judgement needs to be studied.

Although certainly not flawless, the cross comprehension test has its own advantage. A simple instruction set is effective; it encourages authors to make their best effort to put as much information into a short summary. Most importantly, the test is robust; it sometimes causes ill-framed questions, but they can be compensated by relative comparison achieved by crossing summary-questionnaire pairs.

6 Evaluation of Machine Generated Summaries

The objective of this evaluation is to measure the information content of machine generated summaries using a human authored summary as a yardstick. Although very subjective for many cases, a human summary can still be a reference if we do not treat them as a 'gold standard'.

The cross comprehension test of machine generated and human authored summaries is illustrated in Figure 3. Questions are set by the different author from the one who wrote the summary. A human authored summary may still be the best summary in many respects, but it will no longer be considered perfect. One may target the relevance level of the human summary (*e.g.*, 61% for the 'one line' summary task from the broadcast news stories) for automatic summarisation research.

Machine generated summary: senate to vote to approve the expansion of north atlantic treaty organisation to bigger nato means us obligations **Summary by subject B:** US Senate to decide on NATO expansion; US assesses bigger NATO more arms deal but poor ties with Russia. **Questions by subject D:** 1. What is happening to the NATO? 2. Who sees this move as a threat? 3. Who is bearing the main cost?

Table 4: Evaluation of machine and human authored summaries using questions by the different subject.

Table 4 shows one example from those with which we are currently experimenting. Answers sought by Subject **D** were '*expansion*', '*Russian*', and '*American taxpayers*', respectively. Given this question set, answers are '*relevant*', '*relevant*', and '*not found*' for the summary by Subject **B**, and answers found in the machine generated summary are '*relevant*', '*not found*', and '*not found*', respectively.

7 Conclusion

In this paper, we have presented the issue of human subjectivity when authoring summaries, with regard to producing a simple, robust evaluation of machine generated summaries. Applying the cross comprehension test on human authored 'one line' summaries from broadcast news stories, we gauged the level of subjectivity among four authors. The instruction set was simple, thus there was enough room for subjectivity. However the approach was robust because the test did not use the absolute score, instead relying on relative comparison, effectively alleviating the subjectivity. We also showed the approach to evaluating machine generated summaries. The experiment using this scheme is currently underway.

Acknowledgement. This work was funded by UK EPSRC grant GR/R42405, *Statistical Summarisation of Spoken Language* (S3L).

References

C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel. 1999. The TDT-2 text and speech corpus. *DARPA Broadcast News Workshop*, Herndon, VA.

M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui. 2005. Sentence extraction-based presentation summarization techniques and evaluation metrics. *ICASSP*, Philadelphia.

L. Hirschmann, J. Burger, D. Palmer, and P. Robinson. 1999. Evaluating content extraction from audio source. *ESCA Workshop: Accessing Information in Spoken Audio*, Cambridge.

C. Hori, T. Hori, and S. Furui. 2003. Evaluation method for automatic speech summarization. *Eurospeech*, Geneva.

C. Lin and E. Hovy. 2003a. Automatic evaluation of summaries using n-gram co-occurrence statistics. *HLT-NAACL*, Edmonton.

C. Lin and E. Hovy. 2003b. The potential and limitations of automatic sentence extraction for summarization. *HLT-NAACL Workshop on Automatic Summarization*, Edmonton.

I. Mani. 2001. *Automatic Summarization*. Jon Benjamins Publishing Company.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *HLT-NAACL*, Boston.

P. Over and J. Yen. 2004. An introduction to DUC 2004: Intrinsic evaluation of generic news text summarization systems. *DUC Workshop*, Boston.

D. Radev and D. Tam. 2003. Summarization evaluation via relative utility. *CIKM*, New Orleans.

H. Van Halteren and S. Teufel. 2003. Examining the consensus between human summaries: Initial experiments with factoid analysis. *HLT-NAACL Workshop on Automatic Summarization*, Edmonton.

Appendix

Attached below is a complete news story for the human authored summaries in Section 3. It is taken from '*ABC News*', aired on 13 February 1998, between 1830 and 1900 hours, where Peter Jennings was in the studio, and John Mackenzie was reporting from Central Park:

"In case you had forgotten and probably you shouldn't have tomorrow is Valentine's day. It is largely an American celebration though other parts of the world are picking up on it. We're told whether we are married or single male or female. We know what we're supposed to do don't we? We thought tonight without getting too serious about it we would take a closer look at the science of love. I confess that we never thought of it as science until yesterday when we went across to the Central Park Zoo here in New York to meet and have a conversation with Dr Helen Fisher from Rutgers University in New Jersey. Dr Fisher is a noted anthropologist who has been studying the behaviour between men and women for many years."
"Dr Fisher can I ask you is this really serious science that you do?"
"I think it's serious yes. I'm interested in minding the brain physiology of very basic human mating emotion attraction and I think it comes out of nature."
"Can you break down for me what the components of attraction are?"
"It begins when a person takes on special meaning. Indeed you focus on that individual. There's another thing called intrusive thinking. The person pops into your brain. Some people have said I think about him or her eight five of the day. People also focus on their sweetheart. They will remember a tiny little thing that the person said or did. Just the way they toss their head when they got off the bus or reached for the salt at the dinner table. And then of course there's that elation and that giddiness and euphoria and that tremendous despair when the person doesn't call you. When men and women begin to fall in love do they do it differently. Men tend to fall in love faster. I think because men are more turned on by the visual image. A man can scan a room and see a woman who really appeals to him. The woman has to find out whether the man has he resources whether he's a find individual."
"What are the visual traits for a man when he's in this process?"
"The visual traits that a man will be attracted to will vary dramatically. We all grow up as small from childhood and we build an unconscious love map. A whole list of traits that we are individually looking for in a mate. For example our father's sense of humour. The amount of chaos around the house. Subtle little things will get into the brain and we will create almost a testimony plate of what we are looking for."
"Are different cultures attracted in different ways?"
"There are some ways in which people in every culture are attracted in the same way. Men around the world are attracted to women who give off signs of fertility. Clear skin bright eyes a great personality the kinds of things that indicate that a woman would be good at bearing his young. Women around the world are interested in men who have resources status class money the kinds of things that would help them rear their young. Around the world both men and women are attracted to a face that is symmetrical."
"Doesn't matter whether you are Asian or American?"
"No, you and I could go to New Guinea and you and I would be able to pick out what we regarded as the most beautiful woman in the village and the villagers would agree with us."
"What's the difference of the attraction being dominated by brain and culture?"
"I think human beings evolve certain circuits in the brain that light up when you see the right person. Those circuits are largely associated with dopamine chemicals in the brain that give you that sense of elation and giddiness and euphoria when you see the right person. That's the brain chemistry of romance."
"But who you fall in love with when you fall in love where you fall in love how you express your love that's cultural?"
"That you learn."

Preprocessing and Normalization for Automatic Evaluation of Machine Translation

Gregor Leusch and **Nicola Ueffing** and **David Vilar** and **Hermann Ney**
Lehrstuhl für Informatik VI
RWTH Aachen University
D-52056 Aachen, Germany,
{leusch,ueffing,vilar,ney}@i6.informatik.rwth-aachen.de

Abstract

Evaluation measures for machine translation depend on several common methods, such as preprocessing, tokenization, handling of sentence boundaries, and the choice of a reference length. In this paper, we describe and review some new approaches to them and compare these to state-of-the-art methods. We experimentally look into their impact on four established evaluation measures. For this purpose, we study the correlation between automatic and human evaluation scores on three MT evaluation corpora. These experiments confirm that the tokenization method, the reference length selection scheme, and the use of sentence boundaries we introduce will increase the correlation between automatic and human evaluation scores. We find that ignoring case information and normalizing evaluator scores has a positive effect on the sentence level correlation as well.

1 Introduction

Machine translation (MT), as any other natural language processing (NLP) research subject, depends on the evaluation of its results. Unfortunately, human evaluation of MT system output is a time consuming and expensive task. This is why automatic evaluation is preferred to human evaluation in the research community.

Over the last years, a manifold of automatic evaluation measures has been proposed and studied. This underlines the importance, but also the complexity of finding a suitable evaluation measure for MT. We will give a short overview of some measures in section 2 of this paper.

Although most of these measures share similar ideas and foundation, we observe that researchers tend to approach problems common to several measures differently from each other. A noteworthy example here is the determination of a translation reference length.

In section 3, we will have a look onto structural similarities and differences among several measures, focussing on common steps. We will show that decisions taken about them can be as important to the outcome of an evaluation, as the choice of the evaluation measure itself.

To this end, we will study the performance of each error measure and setting by comparison with human evaluation on three different evaluation tasks in section 4. These experiments will show that sophisticated tokenization as well as adding sentence boundaries and a good choice for the reference lengths will improve correlation between automatic and human evaluation significantly. Case normalization and evaluator normalization are helpful only when evaluating on sentence level; system level evaluation is not affected by these methods.

After a discussion of these results in section 5, we will conclude this paper in section 6.

2 Automatic evaluation measures

The majority of MT evaluation approaches are based on the distance or similarity of MT candidate output to a set of reference translations, i.e. to sentences which are known to be correct. The lower this distance is, or the higher the similarity, the better the

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 17–24, Ann Arbor, June 2005.

candidate translations are considered to be, and thus the better the MT system.

2.1 Evaluation measures studied

Out of the vast amount of measures, we will focus on the following measures that are widely used in research and in evaluation campaigns: WER, PER, BLEU, and NIST.

Let a test set consist of $k = 1, \ldots, K$ candidate sentences E_k generated by an MT system. For each candidate sentence E_k, we have a set of $r = 1, \ldots, R_k$ reference sentences $\widetilde{E}_{r,k}$. Let I_k denote the length, and I_k^* the reference length for each sentence E_k. We will explain in section 3.3 how the reference length is calculated.

With this, we write the total candidate length over the corpus as $\bar{I} := \sum_k I_k$, and the total reference length as $\bar{I}^* := \sum_k I_k^*$.

Let $n_{e_1^m,k}$ denote the count of the m-gram e_1^m within the candidate sentence E_k; similarly let $\tilde{n}_{e_1^m,r,k}$ denote the same count within the reference sentence $\widetilde{E}_{r,k}$. The total m-gram count over the corpus is then $\bar{n}_m := \sum_k \sum_{e_1^m \in E_k} n_{e_1^m,k}$.

2.1.1 WER

The word error rate is defined as the Levenshtein distance $d_L(E_k, \widetilde{E}_{r,k})$ between a candidate sentence E_k and a reference sentence $\widetilde{E}_{r,k}$, divided by the reference length I_k^* for normalization.

For a whole candidate corpus with multiple references, we define the WER to be:

$$\text{WER} := \frac{1}{\bar{I}^*} \sum_k \min_r d_L\big(E_k, \widetilde{E}_{r,k}\big)$$

Note that the WER of a single sentence can be calculated as the WER for a corpus of size $K = 1$.

2.1.2 PER

The position independent error rate (Tillmann et al., 1997) ignores the ordering of the words within a sentence. Independent of the word position, the minimum number of deletions, insertions, and substitutions to transform the candidate sentence into the reference sentence is calculated. Using the counts $n_{e,r}$, $\tilde{n}_{e,r,k}$ of a word e in the candidate sentence E_k, and the reference sentence $\widetilde{E}_{r,k}$, we can calculate this distance as

$$d_{\text{PER}}\big(E_k, \widetilde{E}_{r,k}\big) := \frac{1}{2}\Big(\big|I_k - \tilde{I}_k\big| + \sum_e \big|n_{e,k} - \tilde{n}_{e,r,k}\big|\Big)$$

This distance is then normalized into an error rate, the PER, as described in section 2.1.1.

A promising approach is to compare bigram or arbitrary m-gram count vectors instead of unigram count vectors only. This will take into account the ordering of the words within a sentence implicitly, although not as strong as the WER does.

2.1.3 BLEU

BLEU (Papineni et al., 2001) is a precision measure based on m-gram count vectors. The precision is modified such that multiple references are combined into a single m-gram count vector, $\tilde{n}_{e,k} := \max_r \tilde{n}_{e,r,k}$. Multiple occurrences of an m-gram in the candidate sentence are counted as correct only up to the maximum occurrence count within the reference sentences. Typically, $m = 1, \ldots, 4$.

To avoid a bias towards short candidate sentences consisting of "safe guesses" only, sentences shorter than the reference length will be penalized with a brevity penalty.

$$\text{BLEU} := lp_{\text{BLEU}} \cdot \underset{m}{gm} \left\{ \frac{1}{s_m + \bar{n}_m} \cdot \Big(s_m + \sum_k \sum_{e_1^m \in E_k} \min\big(n_{e_1^m,k}\,,\, \tilde{n}_{e_1^m,k}\big)\Big) \right\}$$

with the geometric mean gm and a brevity penalty

$$lp_{\text{BLEU}} := \min\left(1\,,\, \exp\Big(1 - \frac{\bar{I}^*}{\bar{I}}\Big)\right)$$

In the original BLEU definition, the smoothing term s_m is zero. To allow for sentence-wise evaluation, Lin and Och (2004) define the BLEU-S measure with $s_1 := 1$ and $s_{m>1} := 0$. We have adopted this technique for this study.

2.1.4 NIST

The NIST score (Doddington, 2002) extends the BLEU score by taking information weights of the m-grams into account. The NIST information weight is defined as

$$\mathit{Info}(e_1^m) := -\big(\log_2 \bar{\tilde{n}}_{e_1^m} - \log_2 \bar{\tilde{n}}_{e_1^{m-1}}\big)$$

with $\bar{\tilde{n}}_{e_1^m} := \sum_{k,r} \tilde{n}_{e_1^n,k,r}$.

Note that the weight of a phrase occurring in many references sentence for a candidate is considered to be lower than the weight of a phrase occurring only once!

The NIST score is the sum over all information counts of the co-occurring m-grams, summed up separately for each $m = 1, \ldots, 5$ and normalized by the total m-gram count.

$$\text{NIST} := lp_{\text{NIST}} \cdot \sum_m \left(\frac{1}{\bar{n}_m} \cdot \sum_k \sum_{e_1^m \in E_k} \min\left(n_{e_1^m,k}\,,\, \tilde{n}_{e_1^m,k}\right) \cdot \mathit{Info}(e_1^m) \right)$$

As in BLEU, there is a brevity penalty to avoid a bias towards short candidates:

$$lp_{\text{NIST}} := \exp\left(\beta \cdot \log_2^2 \min\left(1\,,\, \frac{\bar{I}}{\bar{I}^*}\right)\right)$$

$$\text{where } \beta := -\frac{\log_2 2}{\log_2^2 3}$$

Due to the information weights, the value of the NIST score depends highly on the selection of the reference corpus. This must be taken into account when comparing NIST scores of different evaluation campaigns.

2.2 Other measures

Lin and Och (2004) introduce a family of three measures named ROUGE. ROUGE-S is a skip-bigram F-measure. ROUGE-L and ROUGE-W are measures based on the length of the longest common subsequence of the sentences. ROUGE-S has a structure similar to the bigram PER presented here. We expect ROUGE-L and ROUGE-W to have similar properties to WER.

In (Leusch et al., 2003), we have described INVWER, a word error rate enhanced by block transposition edit operations. As structure and scores of INVWER are similar to WER, we have omitted INVWER experiments in this paper.

3 Preprocessing and normalization

Although the general idea is clear, there are still several details to be specified when implementing and using an automatic evaluation measure. We are going to investigate the following problems:

The first detail we have to state more precisely is the term "word" in the above formulae. A common approach for western languages is to consider spaces as separators of words. The role of punctuation marks in tokenization is arguable though. A punctuation mark can separate words, it can be part of a word, and it can be a word of its own. Equally it can be irrelevant at all for evaluation.

On the same lines it is to be specified whether we consider words to be equal if they differ only with respect to upper and lower case. For the IWSLT evaluation, (Paul et al., 2004) give an introduction to how the handling of punctuation and case information may affect automatic MT evaluation.

Also, a method to calculate the "reference length" must specified if there are multiple reference sentences of different length.

Since we want to compare automatic evaluation with human evaluation, we have to clarify some questions about assessing human evaluation as well: Large evaluation tasks are usually distributed to several human evaluators. To smooth evaluation noise, it is common practice to have each candidate sentence evaluated by at least two human judges independently. Therefore there are several evaluation scores for each candidate sentence. We require a single score for each system, though. Consequently, we have to specify how to combine the evaluator scores into sentence scores and then the sentence scores into a system score.

Different definitions of this will have a significant impact on automatic and human evaluation scores.

3.1 Tokenization and punctuation

The importance of punctuation as well as the strictness of punctuation rules depends on the language. In most western languages, correct punctuation can vastly improve the legibility of texts. Marks like full stop or comma separate words. Other marks like apostrophes and hyphens can be used to join words, forming new words by this. For example, the spelling "There's" is a contraction of "There is".

Similar phenomena can be found in other languages, although the set of critical characters may vary. Even when evaluating English translations, the candidate sentences may contain source language parts like proper names which should thus be treated according to the source language.

From the viewpoint of an automatic evaluation measure, we have to decide which units we would consider to be words of their own.

We have studied four tokenization methods. The simplest method is keeping the original sentences, and considering only spaces as word separators. Moreover, we can consider all punctuation marks to separate words but remove them completely then. The `mteval` tool (Papineni, 2002) improves this

Table 1: Tokenization methods studied

- Original candidate
  `Powell said: "We'd not be alone; that's for sure."`
- Remove punctuation
  `Powell said We d not be alone that s for sure`
- Tokenization of punctuation (`mteval`)
  `Powell said : " We'd not be alone ; that's for sure . "`
- Tokenization and treatment of abbreviations and contractions
  `Powell said : " we would not be alone ; that is for sure . "`

scheme by keeping all punctuation marks as separate words except for decimal points and hyphens joining composita. We have extended this scheme by implementing a treatment of common English contractions. Table 1 illustrates these methods.

3.2 Case sensitivity

In western languages, maintaining correct upper and lower case can improve the readability of a text. Unfortunately, though the case of a word depends on the word class, classification is not always unambiguous. What is more, the first word in a sentence is always written in upper case. This lowers the significance of case information in MT evaluation, as even a valid reordering of words between candidate and reference sentence may lead to conflicting cases. Consequently, we investigated if and how case information can be exploited for automatic evaluation.

3.3 Reference length

Each automatic evaluation measure we have taken into account depends on the calculation of a reference length: WER, PER, and ROUGE are normalized by it, whereas NIST or BLEU incorporate it for the determination of the brevity penalty. In MT evaluation practise, there are multiple reference sentences for each candidate sentence, with different lengths each. It is thus not intuitively clear what the "reference length" is.

A simple choice here is the average length of the reference sentences. Though this is modus operandi for NIST, it is problematic with brevity penalty or F-measure based scores, as even candidate sentences that are identical to a shorter-than-average reference sentence – which we would intuitively consider to be "optimal" – will then receive a sub-optimal score.

BLEU incorporates a different method for the determination of the reference length in its default implementation: Reference length here is the reference sentence length which is closest to the candidate length. If there is more than one the shortest of them is chosen.

For measures based on the comparison of single sentences such as WER, PER, and ROUGE, at least two more methods deserve consideration:

- The average length of the sentences with the lowest absolute distance or highest similarity to the candidate sentence. We call this method "average nearest-sentence length".
- The length of the sentence with the lowest relative error rate or the highest relative similarity. We call this method "best length". Note that when using this method, not the minimum absolute distance is used for the error rate, but the distance that leads to minimum relative error.

Other strategies studied by us, e.g. minimum length of the reference sentences, did not show any theoretical or experimental advantage over the methods mentioned here. Thus we will not discuss them in this paper.

3.4 Sentence boundaries

The position of a word within a sentence can be quite significant for the correctness of the sentence.

WER, INVWER, and ROUGE-L take into account the ordering explicitly. This is not the case with n-PER, BLEU, or NIST, although the positions of inner words are regarded implicitly by m-gram overlap. To model the position of words at the initial or the end of a sentence, one can enclose the sentence with artificial sentence boundary words. Although this is a common approach in language modelling, it has to our knowledge not yet been applied to MT evaluation.

3.5 Evaluator normalization

For human evaluation, it has to be specified how to handle evaluator bias, and how to combine sentence scores into system scores.

Regarding evaluator bias, even accurate evaluation guidelines will not prevent a measurable discrepancy between the scores assigned by different human evaluators.

The 2003 TIDES/MT evaluation may serve as an example here: Since the candidate sentences of

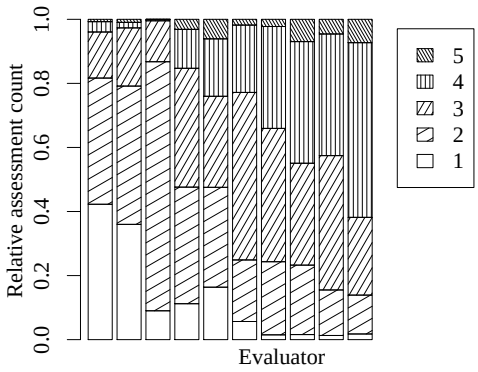


Figure 1: Distribution of adequacy assessments for each human evaluator. TIDES CE corpus.

the participating systems were randomly distributed among ten human evaluators, one would expect the assessed scores to be independent of the evaluator. Figure 1 indicates that this is indeed not the case, as the evaluators can clearly be distinguished by the amount of good and bad marks they assessed.

$(0, 1)$ evaluator normalization overcomes this bias: For each human evaluator the average sentence score given by him or her and its variance are calculated. These assignments are then normalized to $(0, 1)$ expectation and standard deviation (Doddington, 2003), separately for each evaluator.

Evaluator normalization should be unnecessary for system evaluation, as the evaluator biases tend to cancel out over the large amount of candidate sentences if the alignment of evaluators and systems is random enough. Moreover, with $(0, 1)$ normalization the calculated system scores are relative, not absolute scores. As such they can only be compared with scores out of the same evaluation.

Whereas the assessments by the human evaluators are given on the sentence level, our interest may lie on the evaluation of whole candidate systems. Depending on the number of assessments per candidate sentence, different combination methods for the sentence scores can be considered for this, e.g. mean or median. As our data consisted only of two or three human assessments per sentence, we have only applied the mean in our experiments.

It has to be defined how a system score is calculated from the sentence scores. All of the automatic evaluation measures implicitly weight the candidate sentences by their length. Consequently, we applied for the human evaluation scores a weighting by length on sentence level as well.

Table 2: Corpus statistics

	TIDES CE	TIDES AE	BTEC CE
Source language	Chinese	Arabic	Chinese
Target language	English	English	English
Sentences	919	663	500
Running words	25784	17763	3632
Punctuation marks	3760	2698	610
Ref. translations	4	4	16
Avg. ref. length	28.1	26.8	7.3
Candidate systems	7	6	11

4 Experimental results

To assess the impact of the mentioned preprocessing steps, we calculated scores for several automatic evaluation measures with varying preprocessing, reference length calculation, etc. on three evaluation test sets from international MT evaluation campaigns. We then compared these automatic evaluation results with human evaluation of adequacy and fluency by determining a correlation coefficient between human and automatic evaluation. We chose Pearson's r for this. Although all evaluation measures were calculated using length weighting, we did not do any weighting when calculating the sentence level correlation.

Regarding the m-gram PER, we had studied m-gram lengths of up to 8 both separately and in combination with shorter m-gram lengths in previous experiments. However, an m-gram length of greater than 4 did not show noteworthy correlation. For this, we will leave out these results in this paper.

For the sake of clarity, we will also leave out measures that behave very similarly to akin measures e.g. INVWER and WER, 2-PER and 1-PER, or BLEU and BLEU-S.

Since WER and PER are error measures, whereas BLEU and NIST are similarity measures, the correlation coefficients with human evaluation will have opposite signs. For convenience, we will look at the absolute coefficients only.

4.1 Corpora

From the 2003 TIDES evaluation campaign we included both the Chinese-English and the Arabic-English test corpus in our experiments. Both were provided with adequacy and fluency scores between 1 and 5 for seven and six candidate sets respectively.

As we wanted to perform experiments on a corpus with a larger amount of MT systems, we also included the IWSLT BTEC 2004 Chinese-English

evaluation (Akiba et al., 2004). We restricted our experiments to the eleven MT systems that had been trained on a common training corpus.

Corpus statistics can be found in table 2.

4.2 Experimental baseline

In our first experiment we studied the correlation of the different evaluation measures with human evaluation at "baseline" conditions. These included no sentence boundaries, but tokenization with treatment of abbreviations, see table 1. For sentence evaluation, conditions included evaluator normalization. Case information was removed. We used these settings in the other experiments, too, if not stated otherwise.

Figure 2 shows the correlation between automatic and human scores. On the TIDES corpora the system level correlation is particularly high, at a moderate sentence level correlation. We assume the latter is due to the poor sentence inter-annotator agreement on these corpora, which is then smoothed out on system level. On the BTEC corpus a high sentence level correlation accompanies a significantly lower system level correlation. Note that due to the much lower number of samples on the system level (e.g. 5 vs. 5500), small changes in the sentence level correlation are more likely to be significant than such changes on system level. We have verified these effects by inspecting the rank correlation on both levels, as well as by experiments on other corpora. Although these experiments support our findings, we have omitted results here for the sake of clarity.

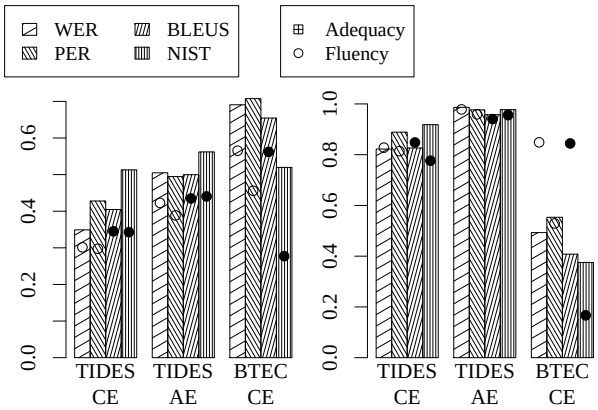


Figure 2: Pearson correlation coefficient between automatic and human evaluation. Bars indicate correlation with adequacy, circles with fluency score.
Left: sentence, **right:** system level correlation.

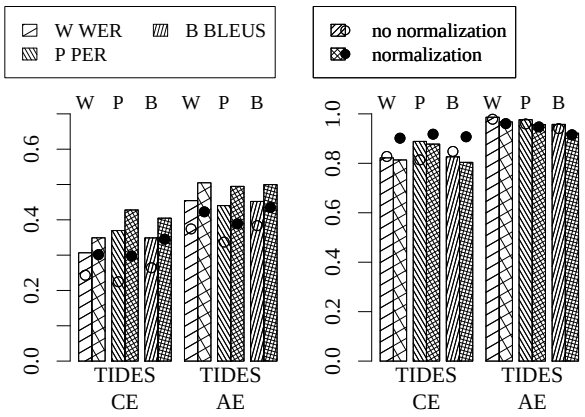


Figure 3: Effect of evaluator normalization.
Left: sentence, **right:** system level correlation.

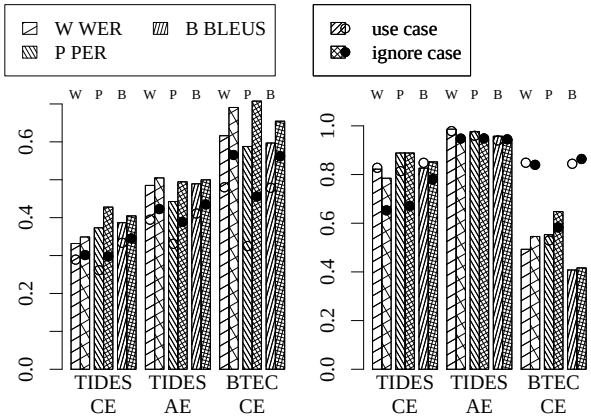


Figure 4: Effect of case normalization.
Left: sentence, **right:** system level correlation.

4.3 Evaluator normalization

We studied the effect of $(0, 1)$-normalization of scores assigned by human evaluators. The NIST measure showed a behavior very similar to that of the other measures and is thus left out in the graph. The correlation of all automatic measures both with fluency and with adequacy increases significantly at sentence level (figure 3). We do not notice a positive effect on system level, which confirms the assumption stated in section 3.5.

4.4 Tokenization and case normalization

The impact of case information was analyzed in our next experiment. Figure 4 (again without the NIST measure as it shows a similar behavior to the other measures) indicates that it is advisable to disregard case information when looking into adequacy on sentence level. Surprisingly, this also holds for

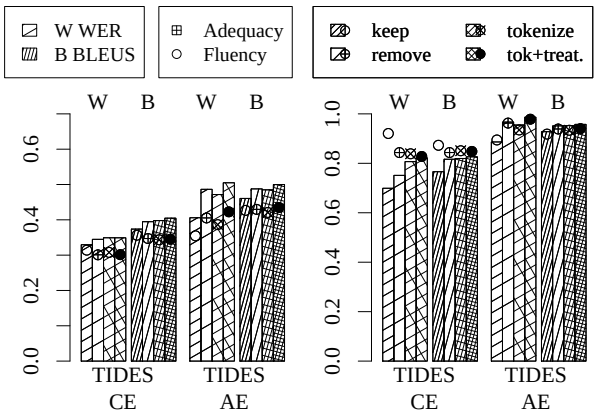


Figure 5: Effect of different tokenization steps.
Left: sentence, **right:** system level correlation.

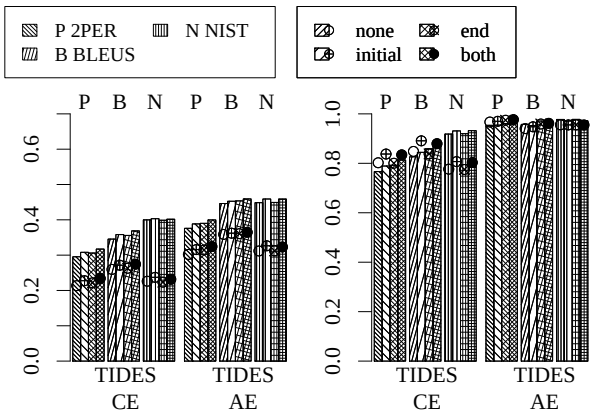


Figure 7: Effect of sentence boundaries.
Left: sentence, **right:** system level correlation.

fluency. We do no find a clear tendency on whether or not to regard case information at system level.

Figure 5 indicates that the way of handling punctuation we proposed does pay off when evaluating adequacy. For fluency our results were contradictory: A slight decrease on the Arabic-English corpus is accompanied by a slight decay on the Chinese-English corpus. We did not investigate the BTEC corpus here as most systems sticked to the tokenization guidelines for this evaluation.

4.5 Reference length

The dependency of evaluation measures on the selection of reference lengths is rarely covered in the literature. However, as we can see in figure 6, our experiments indicate a significant impact. The selected three methods here are the default for WER/PER, NIST, and BLEU, respectively. For the distance based evaluation measures, represented by WER here, taking the length of the sentence leading to the best score leads to the best correlation with both fluency and adequacy. Taking the average length instead seems to be the worst choice.

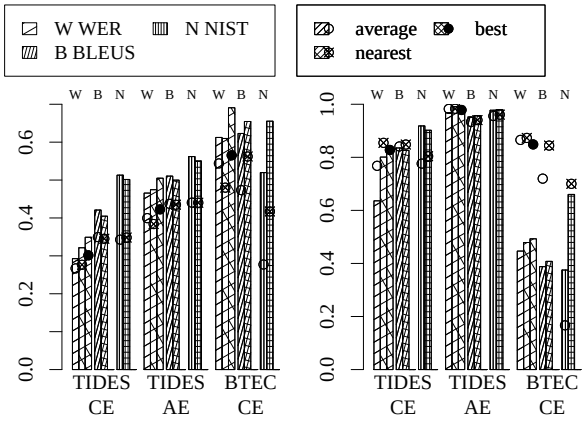


Figure 6: Effect of different reference lengths.
Left: sentence, **right:** system level correlation.

For brevity penalty based measures, the effect is not as clear: On both TIDES corpora there is no significant difference in correlation between using the average length and the nearest length. On the BTEC corpus, choosing the nearest sentence length leads to a significantly higher correlation than choosing the average length. We assume this is due to the high number of reference sentences on this corpus.

4.6 Sentence boundaries

As sentence boundaries will only influence m-gram count vector based measures, we have restricted our experiments to bigram PER, BLEU-S, and NIST here. Including sentence boundaries (figure 7) has a positive effect on correlation with fluency and adequacy for both bigram PER and BLEU-S. Sentence initials seem to be more important than sentence ends here. For the NIST measure, we do not find any significant effect.

5 Discussion

In a perfect MT world, any dependency of an evaluation on case information or tokenization should be inexistent, as MT systems already have to deal with both in the translation process, and could be designed to produce output according to evaluation campaign guidelines. Once all translation systems stick to the same specifications, no further preprocessing steps should be necessary.

In practice there will be some systems that step

out of line. If we then choose strict rules regarding case information and punctuation, automatic error measures will penalize these systems rather hard, whereas penalty is rather low if we choose lax ones.

In this situation case information will have a large effect on the correlation between automatic and human evaluation, depending on whether the involved candidate systems will have a good or a bad human evaluation. It is vital to keep this in mind when drawing conclusions here regarding system evaluation, despite the obvious importance of case information in natural languages.

These considerations also hold for the treatment of punctuation marks, as a special care should be unnecessary if all systems sticked to tokenization specifications. In practise, MT systems differ in the way they generate and handle punctuation marks. Therefore, appropriate preprocessing steps are advisable.

Our experiments suggest that sentence boundaries increase correlation between automatic scores and adequacy both on sentence and on system level. For fluency, the improvement is less significant, and mainly depends on the sentence initials.

For length penalty based measures, we have found that choosing the nearest sentence length yields the highest correlation with human evaluation. For distance based measures instead, it seems advisable to choose the sentence that leads to the best relative score as the one that determines the reference length.

6 Conclusion

We have described several MT evaluation measures. We have pointed out common preprocessing steps and auxiliary methods which have not been studied in detail so far in spite of their importance for the MT evaluation process. Particularly, we have introduced a novel method for determining the reference length of an evaluation candidate sentence, and a simple method to incorporate sentence boundary information to m-gram based evaluation measures.

We then have performed several experiments on these methods on three evaluation corpora. The results indicate that both our new reference length algorithm and the use of sentence boundaries improve the correlation of the studied automatic evaluation measures with human evaluation. Furthermore, we have learned that case information should be removed when performing automatic sentence evaluation. On sentence level, evaluator normalization can improve the correlation between automatic and human evaluation.

Acknowledgements

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistische Textübersetzung" (Ne572/5) and by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

References

Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii. 2004. Overview of the IWSLT04 evaluation campaign. In *Proc. IWSLT*, pp. 1–12, Kyoto, Japan, September.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*.

G. Doddington. 2003. NIST MT Evaluation Workshop. Personal communication, July.

G. Leusch, N. Ueffing, and H. Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proc. MT Summit IX*, pp. 240–247, New Orleans, LA, September.

C. Y. Lin and F. J. Och. 2004. Orange: a method for evaluation automatic evaluation metrics for machine translation. In *Proc. COLING 2004*, pp. 501–507, Geneva, Switzerland, August.

K. A. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, September.

K. A. Papineni. 2002. The NIST `mteval` scoring software. `http://www.itl.nist.gov/iad/894.01/tests/mt/resources/scoring.htm`.

M. Paul, H. Nakaiwa, and M. Federico. 2004. Towards innovative evaluation methodologies for speech translation. In *Working Notes of the NTCIR-4 Meeting*, volume 2, pp. 17–21.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based search for statistical translation. In *European Conf. on Speech Communication and Technology*, pp. 2667–2670, Rhodes, Greece, September.

Syntactic Features for Evaluation of Machine Translation

Ding Liu and **Daniel Gildea**
Department of Computer Science
University of Rochester
Rochester, NY 14627

Abstract

Automatic evaluation of machine translation, based on computing n-gram similarity between system output and human reference translations, has revolutionized the development of MT systems. We explore the use of syntactic information, including constituent labels and head-modifier dependencies, in computing similarity between output and reference. Our results show that adding syntactic information to the evaluation metric improves both sentence-level and corpus-level correlation with human judgments.

1 Introduction

Evaluation has long been a stumbling block in the development of machine translation systems, due to the simple fact that there are many correct translations for a given sentence. Human evaluation of system output is costly in both time and money, leading to the rise of automatic evaluation metrics in recent years. The most commonly used automatic evaluation metrics, BLEU (Papineni et al., 2002) and NIST (Doddington, 2002), are based on the assumption that "The closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002). For every hypothesis, BLEU computes the fraction of n-grams which also appear in the reference sentences, as well as a brevity penalty. NIST uses a similar strategy to BLEU but further considers that n-grams with different frequency should be treated differently in the evaluation. It introduces the notion of information weights, which indicate that rarely occurring n-grams count more than those frequently occurring ones in the evaluation (Doddington, 2002). BLEU and NIST have been shown to correlate closely with human judgments in ranking MT systems with different qualities (Papineni et al., 2002; Doddington, 2002).

In the 2003 Johns Hopkins Workshop on Speech and Language Engineering, experiments on MT evaluation showed that BLEU and NIST do not correlate well with human judgments at the sentence level, even when they correlate well over large test sets (Blatz et al., 2003). Kulesza and Shieber (2004) use a machine learning approach to improve the correlation at the sentence level. Their method, based on the assumption that higher classification accuracy in discriminating human- from machine-generated translations will yield closer correlation with human judgments, uses support vector machine (SVM) based learning to weight multiple metrics such as BLEU, NIST, and WER (minimal word error rate). The SVM is trained for differentiating the MT hypothesis and the professional human translations, and then the distance from the hypothesis's metric vector to the hyper-plane of the trained SVM is taken as the final score for the hypothesis.

While the machine learning approach improves correlation with human judgments, all the metrics discussed are based on the same type of information: n-gram subsequences of the hypothesis translations. This type of feature cannot capture the grammaticality of the sentence, in part because they do not take into account sentence-level information. For example, a sentence can achieve an excellent BLEU score without containing a verb. As MT systems improve, the shortcomings of n-gram based evaluation are becoming more apparent. State-of-the-art MT output

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 25–32, Ann Arbor, June 2005.

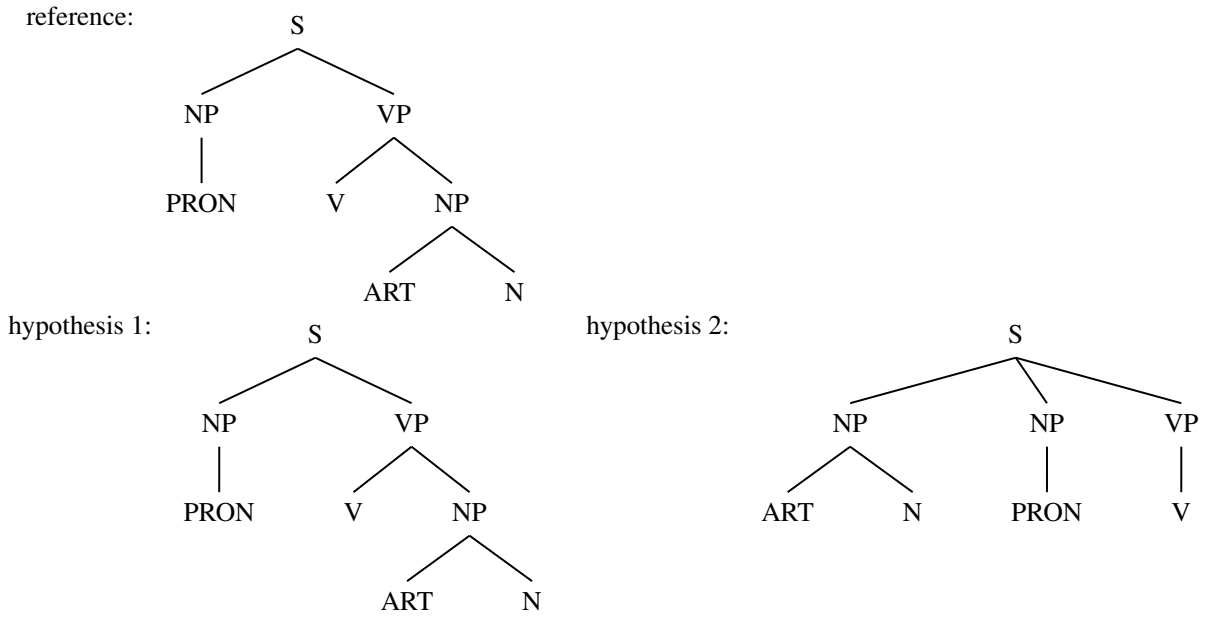


Figure 1: Syntax Trees of the Examples

often contains roughly the correct words and concepts, but does not form a coherent sentence. Often the intended meaning can be inferred; often it cannot. Evidence that we are reaching the limits of n-gram based evaluation was provided by Charniak et al. (2003), who found that a syntax-based language model improved the fluency and semantic accuracy of their system, but lowered their BLEU score.

With the progress of MT research in recent years, we are not satisfied with the getting correct words in the translations; we also expect them to be well-formed and more readable. This presents new challenges to MT evaluation. As discussed above, the existing word-based metrics can not give a clear evaluation for the hypothesis' fluency. For example, in the BLEU metric, the overlapping fractions of n-grams with more than one word are considered as a kind of metric for the fluency of the hypothesis. Consider the following simple example:

Reference: I had a dog.
Hypothesis 1: I have the dog.
Hypothesis 2: A dog I had.

If we use BLEU to evaluate the two sentences, hypothesis 2 has two bigrams *a dog* and *I had* which are also found in the reference, and hypothesis 1 has no bigrams in common with the reference. Thus hypothesis 2 will get a higher score than hypothesis 1. The result is obviously incorrect. However, if we evaluate their fluency based on the syntactic similarity with the reference, we will get our desired results. Figure 1 shows syntactic trees for the example sentences, from which we can see that hypothesis 1 has exactly the same syntactic structure with the reference, while hypothesis 2 has a very different one. Thus the evaluation of fluency can be transformed as computing the syntactic similarity of the hypothesis and the references.

This paper develops a number of syntactically motivated evaluation metrics computed by automatically parsing both reference and hypothesis sentences. Our experiments measure how well these metrics correlate with human judgments, both for individual sentences and over a large test set translated by MT systems of varying quality.

2 Evaluating Machine Translation with Syntactic Features

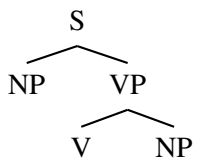
In order to give a clear and direct evaluation for the fluency of a sentence, syntax trees are used to generate metrics based on the similarity of the MT hypothesis's tree and those of the references. We can't expect that the whole syntax tree of the hypothesis can always be found in the references, thus our approach is to be based on the fractions of the subtrees

which also appear in the reference syntax trees. This idea is intuitively derived from BLEU, but with the consideration of the sparse subtrees which lead to zero fractions, we average the fractions in the arithmetic mean, instead of the geometric mean used in BLEU. Then for each hypothesis, the fractions of subtrees with different depths are calculated and their arithmetic mean is computed as the syntax tree based metric, which we denote as "subtree metric" STM:

$$\mathrm{STM} = \frac{1}{D}\sum_{n=1}^{D}\frac{\sum_{t\in \mathrm{subtrees}_n(hyp)} \mathrm{count}_{\mathrm{clip}}(t)}{\sum_{t\in \mathrm{subtrees}_n(hyp)} \mathrm{count}(t)}$$

where D is the maximum depth of subtrees considered, $\mathrm{count}(t)$ denotes the number of times subtree t appears in the candidate's syntax tree, and $\mathrm{count}_{\mathrm{clip}}(t)$ denotes the clipped number of times t appears in the references' syntax trees. Clipped here means that, for a given subtree, the count computed from the hypothesis syntax tree can not exceed the maximum number of times the subtree occurs in any single reference's syntax tree. A simple example with one hypothesis and one reference is shown in Figure 2. Setting the maximum depth to 3, we go through the hypothesis syntax tree and compute the fraction of subtrees with different depths. For the 1-depth subtrees, we get *S*, *NP*, *VP*, *PRON*, *V*, *NP* which also appear in the reference syntax tree. Since *PRON* only occurs once in the reference, its clipped count should be 1 rather than 2. Then we get 6 out of 7 for the 1-depth subtrees. For the 2-depth subtrees, we get *S→NP VP*, *NP→PRON*, and *VP→V NP* which also appear in the reference syntax tree. For the same reason, the subtree *NP→PRON* can only be counted once. Then we get 3 out of 4 for the 2-depth subtree. Similarly, the fraction of 3-depth subtrees is 1 out of 2. Therefore, the final score of STM is (6/7+3/4+1/2)/3=0.702.

While the subtree overlap metric defined above considers only subtrees of a fixed depth, subtrees of other configurations may be important for discriminating good hypotheses. For example, we may want to look for the subtree:



to find sentences with transitive verbs, while ignoring the internal structure of the subject noun phrase. In order to include subtrees of all configurations in our metric, we turn to convolution kernels on our trees. Using $H(x)$ to denote the vector of counts of all subtrees found in tree x, for two trees T_1 and T_2, the inner product $H(T_1)\cdot H(T_2)$ counts the number of matching pairs of subtrees of T_1 and T_2. Collins and Duffy (2001) describe a method for efficiently computing this dot product without explicitly computing the vectors H, which have dimensionality exponential in the size of the original tree. In order to derive a similarity measure ranging from zero to one, we use the cosine of the vectors H:

$$\cos(T_1, T_2) = \frac{H(T_1)\cdot H(T_2)}{|H(T_1)||H(T_2)|}$$

Using the identity

$$|H(T_1)| = \sqrt{H(T_1)\cdot H(T_1)}$$

we can compute the cosine similarity using the kernel method, without ever computing the entire of vector of counts H. Our kernel-based subtree metric TKM is then defined as the maximum of the cosine measure over the references:

$$\mathrm{TKM} = \max_{t\in ref} \cos(\mathrm{hyp}, t)$$

The advantage of using the tree kernel is that it can capture the similarity of subtrees of different shapes; the weak point is that it can only use the reference trees one by one, while STM can use them simultaneously. The dot product also weights individual features differently than our other measures, which compute overlap in the same way as does BLEU. For example, if the same subtree occurs 10 times in both the hypothesis and the reference, this contributes a term of 100 to the dot product, rather than 10 in the clipped count used by BLEU and by our subtree metric STM.

2.1 Dependency-Based Metrics

Dependency trees consist of trees of head-modifier relations with a word at each node, rather than just at the leaves. Dependency trees were found to correspond better across translation pairs than constituent trees by Fox (2002), and form the basis of the machine translation systems of Alshawi et al. (2000)

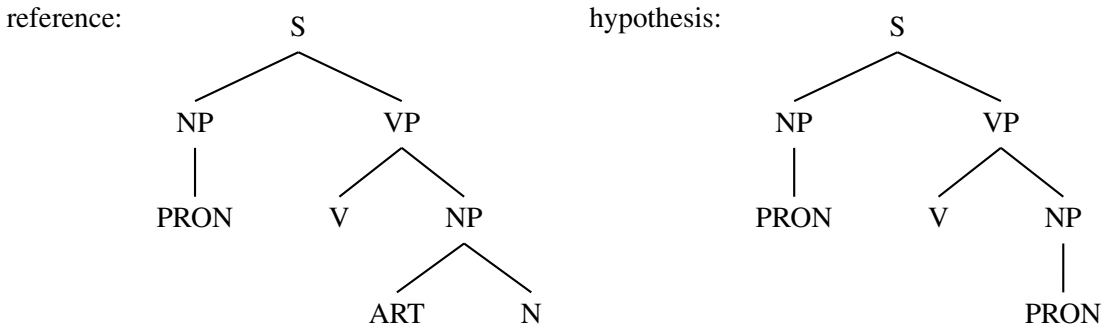
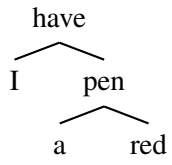


Figure 2: Examples for the Computation of STM

and Lin (2004). We derived dependency trees from the constituent trees by applying the deterministic headword extraction rules used by the parser of Collins (1999). For the example of the reference syntax tree in Figure 2, the whole tree with the root S represents a sentence; and the subtree *NP→ART N* represents a noun phrase. Then for every node in the syntax tree, we can determine its headword by its syntactic structure; from the subtree *NP→ART N*, for example, the headword selection rules chose the headword of *NP* to be word corresponding to the POS *N* in the subtree, and the other child, which corresponds to *ART*, is the modifier for the headword. The dependency tree then is a kind of structure constituted by headwords and every subtree represents the modifier information for its root headword. For example, the dependency tree of the sentence *I have a red pen* is shown as below.



The dependency tree contains both the lexical and syntactic information, which inspires us to use it for the MT evaluation.

Noticing that in a dependent tree the child nodes are the modifier of its parent, we propose a dependency-tree based metric by extracting the headwords chains from both the hypothesis and the reference dependency trees. A headword chain is a sequence of words which corresponds to a path in the dependency tree. Take the dependency tree in Figure 2 as the example, the 2-word headword chains include *have I*, *have pen*, *pen a*, and *pen red*. Before using the headword chains, we need to extract them out of the dependency trees. Figure 3 gives an algorithm which recursively extracts the headword chains in a dependency tree from short to long. Having the headword chains, the headword chain based metric is computed in a manner similar to BLEU, but using n-grams of dependency chains rather than n-grams in the linear order of the sentence. For every hypothesis, the fractions of headword chains which also appear in the reference dependency trees are averaged as the final score. Using HWCM to denote the headword chain based metric, it is computed as follows:

$$\text{HWCM} = \frac{1}{D}\sum_{n=1}^{D}\frac{\sum_{g\in \text{chain}_n(hyp)} \text{count}_{\text{clip}}(g)}{\sum_{g\in \text{chain}_n(hyp)} \text{count}(g)}$$

where D is chosen as the maximum length chain considered.

We may also wish to consider dependency relations over more than two words that are contiguous but not in a single ancestor chain in the dependency tree. For this reason, the two methods described in section 3.1 are used to compute the similarity of dependency trees between the MT hypothesis and its references, and the corresponding metrics are denoted DSTM for dependency subtree metric and DTKM for dependency tree kernel metric.

3 Experiments

Our testing data contains two parts. One part is a set of 665 English sentences generated by a Chinese-English MT system. And for each MT hypothesis, three reference translations are associated with it.

```
Input: dependency tree T, maximum length N of the headword chain
Output: headword chains from length 1 to N

for i = 1 to N
   for every node n in T
      if i == 1
         add n's word to n's 1 word headword chains;
      else
         for every direct child c of n
            for every i-1 words headword chain hc of c
               newchain = joint(n's word, hc);
               add newchain to the i words headword chains of n;
            endfor
         endfor
      endif
   endfor
endfor
```

Figure 3: Algorithm for Extracting the Headword Chains

The human judgments, on a scale of 1 to 5, were collected at the 2003 Johns Hopkins Speech and Language Summer Workshop, which tells the overall quality of the MT hypotheses. The translations were generated by the alignment template system of Och (2003). This testing set is called JHU testing set in this paper. The other set of testing data is from MT evaluation workshop at ACL05. Three sets of human translations (E01, E03, E04) are selected as the references, and the outputs of seven MT systems (E9 E11 E12 E14 E15 E17 E22) are used for testing the performance of our syntactic metrics. Each set of MT translations contains 929 English sentences, each of which is associated with human judgments for its fluency and adequacy. The fluency and adequacy scores both range from 1 to 5.

3.1 Sentence-level Evaluation

Our syntactic metrics are motivated by a desire to better capture grammaticality in MT evaluation, and thus we are most interested in how well they correlate with human judgments of sentences' fluency, rather than the adequacy of the translation. To do this, the syntactic metrics (computed with the Collins (1999) parser) as well as BLEU were used to evaluate hypotheses in the test set from ACL05 MT workshop, which provides both fluency and adequacy scores for each sentence, and their Pearson coefficients of correlation with the human fluency scores were computed. For BLEU and HWCM, in order to avoid assigning zero scores to individual sentences, when precision for n-grams of a particular length is zero we replace it with an epsilon value of 10^{-3}. We choose E14 and E15 as two representative MT systems in the ACL05 MT workshop data set, which have relatively high human scores and low human scores respectively. The results are shown in Table 1 and Table 2, with every metric indexed by the maximum n-gram length or subtree depth. The last row of the each table shows the tree-kernel-based measures, which have no depth parameter to adjust, but implicitly consider all depths.

Max Length/ Depth	BLEU	HWCM	STM	DSTM
1	0.126	0.130	—	—
2	0.132	0.142	0.142	0.159
3	0.117	0.157	0.147	0.150
4	0.093	0.153	0.136	0.121
kernel			0.065	0.090

Table 1: Correlation with Human Fluency Judgments for E14

The results show that in both systems our syntactic metrics all achieve a better performance in the correlation with human judgments of fluency. We also notice that with the increasing of the maximum length of n-grams, the correlation of BLEU with human judgments does not necessarily increase, but decreases in most cases. This is contrary to the argument in BLEU which says that longer n-grams better represent the sentences' fluency than the shorter

Max Length/ Depth	BLEU	HWCM	STM	DSTM
1	0.122	0.128	—	—
2	0.094	0.120	0.134	0.137
3	0.073	0.119	0.144	0.124
4	0.048	0.113	0.143	0.121
kernel			0.089	0.066

Table 2: Correlation with Human Fluency Judgments for E15

Max Length/ Depth	BLEU	HWCM	STM	DSTM
1	0.176	0.191	—	—
2	0.185	0.195	0.171	0.193
3	0.169	0.202	0.168	0.175
4	0.137	0.199	0.158	0.143
kernel			0.093	0.127

Table 3: Correlation with Human Overall Judgments for E14

Max Length/ Depth	BLEU	HWCM	STM	DSTM
1	0.146	0.152	—	—
2	0.124	0.142	0.148	0.152
3	0.095	0.144	0.151	0.139
4	0.067	0.137	0.144	0.137
kernel			0.098	0.084

Table 4: Correlation with Human Overall Judgments for E15

ones. The problem can be explained by the limitation of the reference translations. In our experiments, every hypothesis is evaluated by referring to three human translations. Since the three human translations can only cover a small set of possible translations, with the increasing of n-gram length, more and more correct n-grams might not be found in the references, so that the fraction of longer n-grams turns to be less reliable than the short ones and hurts the final scores. In the the corpus-level evaluation of a MT system, the sparse data problem will be less serious than in the sentence-level evaluation, since the overlapping n-grams of all the sentences and their references will be summed up. So in the traditional BLEU algorithm used for corpus-level evaluation, a maximum n-gram of length 4 or 5 is usually used. A similar trend can be found in syntax tree and dependency tree based metrics, but the decreasing ratios are much lower than BLEU, which indicates that the syntactic metrics are less affected by the sparse data problem. The poor performance of tree-kernel based metrics also confirms our arguments on the sparse data problem, since the kernel measures implicitly consider the overlapping ratios of the sub-trees of all shapes, and thus will be very much affected by the sparse data problem.

Though our syntactic metrics are proposed for evaluating the sentences' fluency, we are curious how well they do in the overall evaluation of sentences. Thus we also computed each metric's correlation with human overall judgments in E14, E15 and JHU testing set. The overall human score for each sentence in E14 and E15 is computed by summing up its fluency score and adequacy score. The results are shown in Table 3, Table 4, and Table 5. We can see that the syntactic metrics achieve competitive correlations in the test, among which HWCM, based on headword chains, gives better performances in evaluation of E14 and E15, and a slightly worse performance in JHU testing set than BLEU. Just as with the fluency evaluation, HWCM and other syntactic metrics present more stable performance as the n-gram's length (subtree's depth) increases.

3.2 Corpus-level Evaluation

While sentence-level evaluation is useful if we are interested in a confidence measure on MT outputs, corpus level evaluation is more useful for comparing

Max Length/ Depth	BLEU	HWCM	STM	DSTM
1	0.536	0.502	—	—
2	0.562	0.555	0.515	0.513
3	0.513	0.538	0.529	0.477
4	0.453	0.510	0.497	0.450
kernel			0.461	0.413

Table 5: Correlation with Human Overall Judgments for JHU Testing Set

Max Length/ Depth	BLEU	HWCM	STM	DSTM
1	0.629	0.723	—	—
2	0.683	0.757	0.538	0.780
3	0.724	0.774	0.597	0.780
4	0.753	0.778	0.612	0.788
5	0.781	0.780	0.618	0.778
6	0.763	0.778	0.618	0.782
kernel			0.539	0.875

Table 6: Corpus-level Correlation with Human Overall Judgments (E9 E11 E12 E14 E15 E17 E22)

MT systems and guiding their development. Does higher sentence-level correlation necessarily indicate higher correlation in corpus-level evaluation? To answer this question, we used our syntactic metrics and BLEU to evaluate all the human-scored MT systems (E9 E11 E12 E14 E15 E17 E22) in the ACL05 MT workshop test set, and computed the correlation with human overall judgments. The human judgments for an MT system are estimated by summing up each sentence's human overall score. Table 6 shows the results indexed by different n-grams and tree depths.

We can see that the corpus-level correlation and the sentence-level correlation don't always correspond. For example, the kernel dependency subtree metric achieves a very good performance in corpus-level evaluation, but it has a poor performance in sentence-level evaluation. Sentence-level correlation reflects the relative qualities of different hypotheses in a MT system, which does not indicate any information for the relative qualities of different systems. If we uniformly decrease or increase every hypothesis's automatic score in a MT system, the sentence-level correlation with human judgments will remain the same, but the corpus-level correlation will be changed. So we might possibly get inconsistent corpus-level and sentence-level correlations.

From the results, we can see that with the increase of n-grams length, the performance of BLEU and HWCM will first increase up to length 5, and then starts decreasing, where the optimal n-gram length of 5 corresponds to our usual setting for BLEU algorithm. This shows that corpus-level evaluation, compared with the sentence-level evaluation, is much less sensitive to the sparse data problem and thus leaves more space for making use of comprehensive evaluation metrics. We speculate this is why the kernel dependency subtree metric achieves the best performance among all the metrics. We can also see that HWCM and DSTM beat BLEU in most cases and exhibit more stable performance.

An example hypothesis which was assigned a high score by HWCM but a low score by BLEU is shown in Table 7. In this particular sentence, the common head-modifier relations "aboard ← plane" and "plane ← the" caused a high headword chain overlap, but did not appear as common n-grams counted by BLEU. The hypothesis is missing the word "fifth", but was nonetheless assigned a high score by human judges. This is probably due to its fluency, which HWCM seems to capture better than BLEU.

4 Conclusion

This paper introduces several syntax-based metrics for the evaluation of MT, which we find to be particularly useful for predicting a hypothesis's *fluency*. The syntactic metrics, except the kernel based ones, all outperform BLEU in sentence-level fluency evaluation. For the overall evaluation of sentences for fluency and adequacy, the metric based on headword chain performs better than BLEU in both sentence-level and corpus-level correlation with human judgments. The kernel based metrics, though poor in sentence-level evaluation, achieve the best results in corpus-level evaluation, where sparse data are less of a barrier.

Our syntax-based measures require the existence of a parser for the language in question, however it is worth noting that a parser is required for the target language only, as all our measures of similarity are defined across hypotheses and references in the same language.

Our results, in particular for the primarily structural STM, may be surprising in light of the fact that the parser is not designed to handle ill-formed or ungrammatical sentences such as those produced by machine translation systems. Modern statistical parsers have been tuned to discriminate good structures from bad rather than good sentences from bad.

hyp	Diplomats will be aboard the plane to return home .
ref1	Diplomats are to come back home aboard the fifth plane .
ref2	Diplomatic staff would go home in a fifth plane .
ref3	Diplomatic staff will take the fifth plane home .

Table 7: An example hypothesis in the ACL05-MTE workshop which was assigned a high score by HWCM (0.511) but a low score by BLEU (0.084). Both human judges assigned a high score (4).

Indeed, in some recent work on re-ranking machine translation hypotheses (Och et al., 2004), parser-produced structures were not found to provide helpful information, as a parser is likely to assign a good-looking structure to even a lousy input hypothesis.

However, there is an important distinction between the use of parsers in re-ranking and evaluation – in the present work we are looking for similarities between pairs of parse trees rather than at features of a single tree. This means that the syntax-based evaluation measures can succeed even when the tree structure for a poor hypothesis looks reasonable on its own, as long as it is sufficiently distinct from the structures used in the references.

We speculate that by discriminatively training weights for the individual subtrees and headword chains used by the syntax-based metrics, further improvements in evaluation accuracy are possible.

Acknowledgments We are very grateful to Alex Kulesza for assistance with the JHU data. This work was partially supported by NSF ITR IIS-09325646 and NSF ITR IIS-0428020.

References

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, 26(1):45–60.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Summer Workshop Final Report.

Eugene Charniak, Kevin Knight, and Kenji Yamada. 2003. Syntax-based language models for machine translation. In *Proc. MT Summit IX*.

Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*.

Michael John Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *In HLT 2002, Human Language Technology Conference*, San Diego, CA.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 304–311.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Baltimore, MD, October.

Dekang Lin. 2004. A path-based transfer model for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 625–630, Geneva, Switzerland.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the 2004 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-04)*, Boston.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41th Annual Conference of the Association for Computational Linguistics (ACL-03)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA.

Evaluating Automatic Summaries of Meeting Recordings

Gabriel Murray
Centre for Speech Technology Research
University of Edinburgh
Edinburgh, United Kingdom

Steve Renals
Centre for Speech Technology Research
University of Edinburgh
Edinburgh, United Kingdom

Jean Carletta
Human Communication Research Centre
University of Edinburgh
Edinburgh, United Kingdom

Johanna Moore
Human Communication Research Centre
University of Edinburgh
Edinburgh, United Kingdom

Abstract

The research below explores schemes for evaluating automatic summaries of business meetings, using the ICSI Meeting Corpus (Janin et al., 2003). Both automatic and subjective evaluations were carried out, with a central interest being whether or not the two types of evaluations correlate with each other. The evaluation metrics were used to compare and contrast differing approaches to automatic summarization, the deterioration of summary quality on ASR output versus manual transcripts, and to determine whether manual extracts are rated significantly higher than automatic extracts.

1 Introduction

In the field of automatic summarization, it is widely agreed upon that more attention needs to be paid to the development of standardized approaches to summarization evaluation. For example, the current incarnation of the Document Understanding Conference is putting its main focus on the development of evaluation schemes, including semi-automatic approaches to evaluation. One semi-automatic approach to evaluation is ROUGE (Lin and Hovy, 2003), which is primarily based on n-gram co-occurrence between automatic and human summaries. A key question of the research contained herein is how well ROUGE correlates with human judgments of summaries within the domain of meeting speech. If it is determined that the two types of evaluations correlate strongly, then ROUGE will likely be a valuable and robust evaluation tool in the development stage of a summarization system, when the cost of frequent human evaluations would be prohibitive.

Three basic approaches to summarization are evaluated and compared below: Maximal Marginal Relevance, Latent Semantic Analysis, and feature-based classification. The other major comparisons in this paper are between summaries on ASR versus manual transcripts, and between manual and automatic extracts. For example, regarding the former, it might be expected that summaries on ASR transcripts would be rated lower than summaries on manual transcripts, due to speech recognition errors. Regarding the comparison of manual and automatic extracts, the manual extracts can be thought of as a gold standard for the extraction task, representing the performance ceiling that the automatic approaches are aiming for.

More detailed descriptions of the summarization approaches and experimental setup can be found in (Murray et al., 2005). That work relied solely on ROUGE as an evaluation metric, and this paper proceeds to investigate whether ROUGE alone is a reliable metric for our summarization domain, by comparing the automatic scores with recently-gathered human evaluations. Also, it should be noted that while we are at the moment only utilizing intrinsic evaluation methods, our ultimate plan is to evaluate these meeting summaries extrinsically within the context of a meeting browser (Wellner et al., 2005).

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 33–40, Ann Arbor, June 2005.

2 Description of the Summarization Approaches

2.1 Maximal Marginal Relevance (MMR)

MMR (Carbonell and Goldstein, 1998) uses the vector-space model of text retrieval and is particularly applicable to query-based and multi-document summarization. The MMR algorithm chooses sentences via a weighted combination of query-relevance and redundancy scores, both derived using cosine similarity. The MMR score $Sc^{MMR}(i)$for a given sentence S_i in the document is given by

$$Sc^{MMR}(i) =$$

$$\lambda(Sim(S_i, D)) - (1-\lambda)(Sim(S_i, Summ))\,,$$

where D is the average document vector, $Summ$ is the average vector from the set of sentences already selected, and λ trades off between relevance and redundancy. Sim is the cosine similarity between two documents.

This implementation of MMR uses lambda annealing so that relevance is emphasized while the summary is still short and minimizing redundancy is prioritized more highly as the summary lengthens.

2.2 Latent Semantic Analysis (LSA)

LSA is a vector-space approach which involves projecting the original term-document matrix to a reduced dimension representation. It is based on the singular value decomposition (SVD) of an $m \times n$ term-document matrix A, whose elements A_{ij} represent the weighted term frequency of term i in document j. In SVD, the term-document matrix is decomposed as follows:

$$A = USV^T$$

where U is an $m \times n$ matrix of left-singular vectors, S is an $n \times n$ diagonal matrix of singular values, and V is the $n \times n$ matrix of right-singular vectors. The rows of V^T may be regarded as defining topics, with the columns representing sentences from the document. Following Gong and Liu (Gong and Liu, 2001), summarization proceeds by choosing, for each row in V^T, the sentence with the highest value. This process continues until the desired summary length is reached.

Two drawbacks of this method are that dimensionality is tied to summary length and that good sentence candidates may not be chosen if they do not "win" in any dimension (Steinberger and Ježek, 2004). The authors in (Steinberger and Ježek, 2004) found one solution, by extracting a single LSA-based sentence score, with variable dimensionality reduction.

We address the same concerns, following the Gong and Liu approach, but rather than extracting the best sentence for each topic, the n best sentences are extracted, with n determined by the corresponding singular values from matrix S. The number of sentences in the summary that will come from the first topic is determined by the percentage that the largest singular value represents out of the sum of all singular values, and so on for each topic. Thus, dimensionality reduction is no longer tied to summary length and more than one sentence per topic can be chosen. Using this method, the level of dimensionality reduction is essentially learned from the data.

2.3 Feature-Based Approaches

Feature-based classification approaches have been widely used in text and speech summarization, with positive results (Kupiec et al., 1995). In this work we combined textual and prosodic features, using Gaussian mixture models for the extracted and non-extracted classes. The prosodic features were the mean and standard deviation of F0, energy, and duration, all estimated and normalized at the word-level, then averaged over the utterance. The two lexical features were both TFIDF-based: the average and the maximum TFIDF score for the utterance.

For our second feature-based approach, we derived single LSA-based sentence scores (Steinberger and Ježek, 2004) to complement the six features described above, to determine whether such an LSA sentence score is beneficial in determining sentence importance. We reduced the original term-document matrix to 300 dimensions; however, Steinberger and Ježek found the greatest success in their work by reducing to a single dimension (Steinberger, personal communication). The LSA sentence score was obtained using:

$$Sc_i^{LSA} = \sqrt{\sum_{k=1}^{n} v(i,k)^2 * \sigma(k)^2}\,,$$

where $v(i,k)$ is the kth element of the ith sentence vector and $\sigma(k)$ is the corresponding singular value.

3 Experimental Setup

We used human summaries of the ICSI Meeting corpus for evaluation and for training the feature-based approaches. An evaluation set of six meetings was defined and multiple human summaries were created for these meetings, with each test meeting having either three or four manual summaries. The remaining meetings were regarded as training data and a single human summary was created for these. Our summaries were created as follows.

Annotators were given access to a graphical user interface (GUI) for browsing an individual meeting that included earlier human annotations: an orthographic transcription time-synchronized with the audio, and a topic segmentation based on a shallow hierarchical decomposition with keyword-based text labels describing each topic segment. The annotators were told to construct a textual summary of the meeting aimed at someone who is interested in the research being carried out, such as a researcher who does similar work elsewhere, using four headings:

- general abstract: "why are they meeting and what do they talk about?";
- decisions made by the group;
- progress and achievements;
- problems described

The annotators were given a 200 word limit for each heading, and told that there must be text for the general abstract, but that the other headings may have null annotations for some meetings.

Immediately after authoring a textual summary, annotators were asked to create an extractive summary, using a different GUI. This GUI showed both their textual summary and the orthographic transcription, without topic segmentation but with one line per dialogue act based on the pre-existing MRDA coding (Shriberg et al., 2004) (The dialogue act categories themselves were not displayed, just the segmentation). Annotators were told to extract dialogue acts that together would convey the information in the textual summary, and could be used to support the correctness of that summary. They were given no specific instructions about the number or percentage of acts to extract or about redundant dialogue act. For each dialogue act extracted, they were then required in a second pass to choose the sentences from the textual summary supported by the dialogue act, creating a many-to-many mapping between the recording and the textual summary.

The MMR and LSA approaches are both unsupervised and do not require labelled training data. For both feature-based approaches, the GMM classifiers were trained on a subset of the training data representing approximately 20 hours of meetings.

We performed summarization using both the human transcripts and speech recognizer output. The speech recognizer output was created using baseline acoustic models created using a training set consisting of 300 hours of conversational telephone speech from the Switchboard and Callhome corpora. The resultant models (cross-word triphones trained on conversational side based cepstral mean normalised PLP features) were then MAP adapted to the meeting domain using the ICSI corpus (Hain et al., 2005). A trigram language model was employed. Fair recognition output for the whole corpus was obtained by dividing the corpus into four parts, and employing a leave one out procedure (training the acoustic and language models on three parts of the corpus and testing on the fourth, rotating to obtain recognition results for the full corpus). This resulted in an average word error rate (WER) of 29.5%. Automatic segmentation into dialogue acts or sentence boundaries was not performed: the dialogue act boundaries for the manual transcripts were mapped on to the speech recognition output.

3.1 Description of the Evaluation Schemes

A particular interest in our research is how automatic measures of informativeness correlate with human judgments on the same criteria. During the development stage of a summarization system it is not feasible to employ many hours of manual evaluations, and so a critical issue is whether or not software packages such as ROUGE are able to measure informativeness in a way that correlates with subjective summarization evaluations.

3.1.1 ROUGE

Gauging informativeness has been the focus of automatic summarization evaluation research. We used the ROUGE evaluation approach (Lin and Hovy, 2003), which is based on n-gram co-occurrence between machine summaries and "ideal" human summaries. ROUGE is currently the standard objective evaluation measure for the Document Understanding Conference [1]; ROUGE does not assume that there is a single "gold standard" summary. Instead it operates by matching the target summary against a set of reference summaries. ROUGE-1 through ROUGE-4 are simple n-gram co-occurrence measures, which check whether each n-gram in the reference summary is contained in the machine summary. ROUGE-L and ROUGE-W are measures of common subsequences shared between two summaries, with ROUGE-W favoring contiguous common subsequences. Lin (Lin and Hovy, 2003) has found that ROUGE-1 and ROUGE-2 correlate well with human judgments.

3.1.2 Human Evalautions

The subjective evaluation portion of our research utilized 5 judges who had little or no familiarity with the content of the ICSI meetings. Each judge evaluated 10 summaries per meeting, for a total of sixty summaries. In order to familiarize themselves with a given meeting, they were provided with a human abstract of the meeting and the full transcript of the meeting with links to the audio. The human judges were instructed to read the abstract, and to consult the full transcript and audio as needed, with the entire familiarization stage not to exceed 20 minutes.

The judges were presented with 12 questions at the end of each summary, and were instructed that upon beginning the questionnaire they should not reconsult the summary itself. 6 of the questions regarded informativeness and 6 involved readability and coherence, though our current research concentrates on the informativeness evaluations. The evaluations used a Likert scale based on agreement or disagreement with statements, such as the following Informativeness statements:

1. The important points of the meeting are represented in the summary.
2. The summary avoids redundancy.
3. The summary sentences on average seem relevant.
4. The relationship between the importance of each topic and the amount of summary space given to that topic seems appropriate.
5. The summary is repetitive.
6. The summary contains unnecessary information.

Statements such as 2 and 5 above are measuring the same impressions, with the polarity of the statements merely reversed, in order to better gauge the reliability of the answers. The readability/coherence portion consisted of the following statements:

1. It is generally easy to tell whom or what is being referred to in the summary.
2. The summary has good continuity, i.e. the sentences seem to join smoothly from one to another.
3. The individual sentences on average are clear and well-formed.
4. The summary seems disjointed.
5. The summary is incoherent.
6. On average, individual sentences are poorly constructed.

It was not possible in this paper to gauge how responses to these readability statements correlate with automatic metrics, for the reason that automatic metrics of readability and coherence have not been widely discussed in the field of summarization. Though subjective evaluations of summaries are often divided into informativeness and readability questions, only automatic metrics of informativeness have been investigated in-depth by the summarization community. We believe that the development of automatic metrics for coherence and readability should be a high priority for researchers in summarization evaluation and plan on pursuing this avenue of research. For example, work on coherence in NLG (Lapata, 2003) could potentially inform summarization evaluation. Mani (Mani et al.,

[1] http://duc.nist.gov/

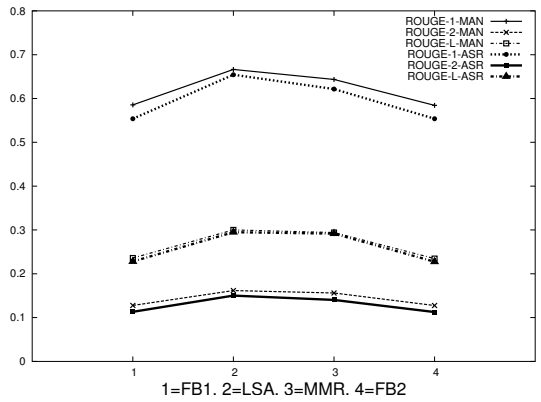


Figure 1: *ROUGE Scores for the Summarization Approaches*

1999) is one of the few papers to have discussed measuring summary readability automatically.

4 Results

The results of these experiments can be analyzed in various ways: significant differences of ROUGE results across summarization approaches, deterioration of ROUGE results on ASR versus manual transcripts, significant differences of human evaluations across summarization approaches, deterioration of human evaluations on ASR versus manual transcripts, and finally, the correlation between ROUGE and human evaluations.

4.1 ROUGE results across summarization approaches

All of the machine summaries were 10% of the original document length, in terms of the number of dialogue acts contained. Of the four approaches to summarization used herein, the latent semantic analysis method performed the best on every meeting tested for every ROUGE measure with the exception of ROUGE-3 and ROUGE-4. This approach was significantly better than either feature-based approach ($p<0.05$), but was not a significant improvement over MMR. For ROUGE-3 and ROUGE-4, none of the summarization approaches were significantly different from each other, owing to data sparsity. Figure 1 gives the ROUGE-1, ROUGE-2 and ROUGE-L results for each of the summarization approaches, on both manual and ASR transcripts.

4.1.1 ASR versus Manual

The results of the four summarization approaches on ASR output were much the same, with LSA and MMR being comparable to each other, and each of them outperforming the feature-based approaches. On ASR output, LSA again consistently performed the best.

Interestingly, though the LSA approach scored higher when using manual transcripts than when using ASR transcripts, the difference was small and insignificant despite the nearly 30% WER of the ASR. All of the summarization approaches showed minimal deterioration when used on ASR output as compared to manual transcripts, but the LSA approach seemed particularly resilient, as evidenced by Figure 1. One reason for the relatively small impact of ASR output on summarization results is that for each of the 6 meetings, the WER of the summaries was lower than the WER of the meeting as a whole. Similarly, Valenza et al (Valenza et al., 1999) and Zechner and Waibel (Zechner and Waibel, 2000) both observed that the WER of extracted summaries was significantly lower than the overall WER in the case of broadcast news. The table below demonstrates the discrepancy between summary WER and meeting WER for the six meetings used in this research.

Meeting	**Summary WER**	**Meeting WER**
Bed004	27.0	35.7
Bed009	28.3	39.8
Bed016	39.6	49.8
Bmr005	23.9	36.1
Bmr019	28.0	36.5
Bro018	25.9	35.6
WER% for Summaries and Meetings		

There was no improvement in the second feature-based approach (adding an LSA sentence score) as compared with the first feature-based approach. The sentence score used here relied on a reduction to 300 dimensions, which may not have been ideal for this data.

The similarity between the MMR and LSA approaches here mirrors Gong and Liu's findings, giving credence to the claim that LSA maximizes relevance and minimizes redundancy, in a different and more opaque manner then MMR, but with similar

STATEMENT	FB1	LSA	MMR	FB2
IMPORT. POINTS	5.03	4.53	4.67	4.83
NO REDUN.	**4.33**	2.60	3.00	3.77
RELEVANT	4.83	4.07	4.33	4.53
TOPIC SPACE	4.43	3.83	3.87	4.30
REPETITIVE	**3.37**	4.70	4.60	3.83
UNNEC. INFO.	**4.70**	6.00	5.83	5.00

Table 1: Human Scores for 4 Approaches on Manual Transcripts

STATEMENT	FB1	LSA	MMR	FB2
IMPORT. POINTS	3.53	**4.13**	3.73	3.50
NO REDUN.	3.40	2.97	2.63	3.57
RELEVANT	3.47	3.57	3.00	3.47
TOPIC SPACE	3.27	3.33	3.00	3.20
REPETITIVE	4.43	4.73	4.70	4.20
UNNEC. INFO.	5.37	6.00	6.00	5.33

Table 2: Human Scores for 4 Approaches on ASR Transcripts

results. Regardless of whether or not the singular vectors of V^T can rightly be thought of as topics or concepts (a seemingly strong claim), the LSA approach was as successful as the more popular MMR algorithm.

4.2 Human results across summarization approaches

Table 1 presents average ratings for the six statements across four summarization approaches on manual transcripts. Interestingly, the first feature-based approach is given the highest marks on each criterion. For statements 2, 5 and 6 FB1 is significantly better than the other approaches. It is particularly surprising that FB1 would score well on statement 2, which concerns redundancy, given that MMR and LSA explicitly aim to reduce redundancy while the feature-based approaches are merely classifying utterances as relevant or not. The second feature-based approach was not significantly worse than the first on this score.

Considering the difficult task of evaluating ten extractive summaries per meeting, we are quite satisfied with the consistency of the human judges. For example, statements that were merely reworded versions of other statements were given consistent ratings. It was also the case that, with the exception of evaluating the sixth statement, judges were able to tell that the manual extracts were superior to the automatic approaches.

4.2.1 ASR versus Manual

Table 2 presents average ratings for the six statements across four summarization approaches on ASR transcripts. The LSA and MMR approaches performed better in terms of having less deterioration of scores when used on ASR output instead of manual transcripts. LSA-ASR was not significantly worse than LSA on any of the 6 ratings. MMR-ASR was significantly worse than MMR on only 3 of the 6. In contrast, FB1-ASR was significantly worse than FB1 for 5 of the 6 approaches, reinforcing the point that MMR and LSA seem to favor extracting utterances with fewer errors. Figures 2, 3 and 4 depict the how the ASR and manual approaches affect the INFORMATIVENESS-1, INFORMATIVENESS-4 and INFORMATIVENESS-6 ratings, respectively. Note that for Figure 6, a higher score is a worse rating.

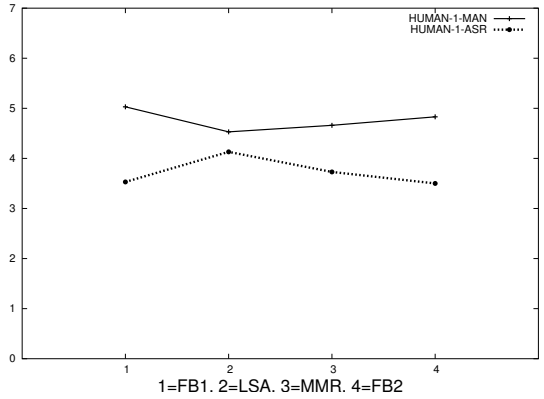


Figure 2: *INFORMATIVENESS-1 Scores for the Summarization Approaches*

4.3 ROUGE and Human correlations

According to (Lin and Hovy, 2003), ROUGE-1 correlates particularly well with human judgments of informativeness. In the human evaluation survey discussed here, the first statement (INFORMATIVENESS-1) would be expected to correlate most highly with ROUGE-1, as it is ask-

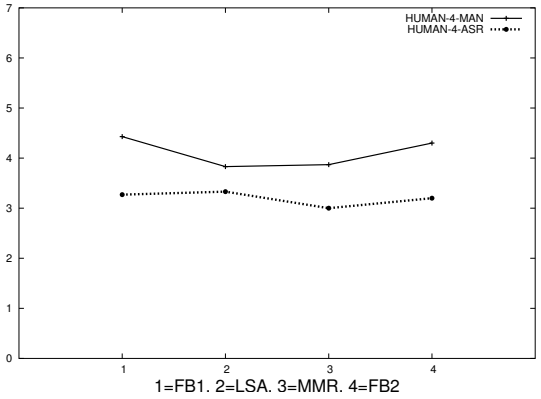


Figure 3: *INFORMATIVENESS-4 Scores for the Summarization Approaches*

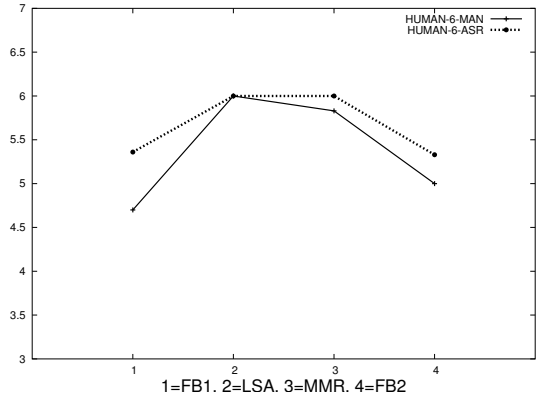


Figure 4: *INFORMATIVENESS-6 Scores for the Summarization Approaches*

ing whether the summary contains the important points of the meeting. As could be guessed from the discussion above, there is no significant correlation between ROUGE-1 and human evaluations when analyzing only the 4 summarization approaches on manual transcripts. However, when looking at the 4 approaches on ASR output, ROUGE-1 and INFORMATIVENESS-1 have a moderate and significant positive correlation (Spearman's rho = 0.500, $p < 0.05$). This correlation on ASR output is strong enough that when ROUGE-1 and INFORMATIVENESS-1 scores are tested for correlation across all 8 summarization approaches, there is a significant positive correlation (Spearman's rho = 0.388, $p < 0.05$).

The other significant correlations for ROUGE-1 across all 8 summarization approaches are with INFORMATIVENESS-2, INFORMATIVENESS-5 and INFORMATIVENESS-6. However, these are negative correlations. For example, with regard to INFORMATIVENESS-2, summaries that are rated as having a high level of redundancy are given high ROUGE-1 scores, and summaries with little redundancy are given low ROUGE-1 scores. Similary, with regard to INFORMATIVENESS-6, summaries that are said to have a great deal of unnecessary information are given high ROUGE-1 scores. It is difficult to interpret some of these negative correlations, as ROUGE does not measure redundancy and would not necessarily be expected to correlate with redundancy evaluations.

5 Discussion

In general, ROUGE did not correlate well with the human evaluations for this data. The MMR and LSA approaches were deemed to be significantly better than the feature-based approaches according to ROUGE, while these findings were reversed according to the human evaluations. An area of agreement, however, is that the LSA-ASR and MMR-ASR approaches have a small and insignificant decline in scores compared with the decline of scores for the feature-based approaches. One of the most interesting findings of this research is that MMR and LSA approaches used on ASR tend to select utterances with fewer ASR errors.

ROUGE has been shown to correlate well with human evaluations in DUC, when used on news corpora, but the summarization task here – using conversational speech from meetings – is quite different from summarizing news articles. ROUGE may simply be less applicable to this domain.

6 Future Work

It remains to be determined through further experimentation by researchers using various corpora whether or not ROUGE truly correlates well with human judgments. The results presented above are mixed in nature, but do not present ROUGE as being sufficient in itself to robustly evaluate a summarization system under development.

We are also interested in developing automatic metrics of coherence and readability. We now have human evaluations of these criteria and are ready to

begin testing for correlations between these subjective judgments and potential automatic metrics.

7 Acknowledgements

Thanks to Thomas Hain and the AMI-ASR group for the speech recognition output. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811, publication).

References

J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. ACM SIGIR*, pages 335–336.

Y. Gong and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. ACM SIGIR*, pages 19–25.

T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, I.Mc.Cowan, J.Vepa, and S.Renals. 2005. An investigation into transcription of conference room meetings. *Submitted to Eurospeech*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proc. IEEE ICASSP*.

J. Kupiec, J. Pederson, and F. Chen. 1995. A trainable document summarizer. In *ACM SIGIR '95*, pages 68–73.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *ACL*, pages 545–552.

C.-Y. Lin and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. HLT-NAACL*.

Inderjeet Mani, Barbara Gates, and Eric Bloedorn. 1999. Improving summaries by revising them. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 558–565, Morristown, NJ, USA. Association for Computational Linguistics.

G. Murray, S. Renals, and J. Carletta. 2005. Extractive summarization of meeting recordings. *Submitted to Eurospeech*.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, , and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100.

J. Steinberger and K. Ježek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proc. ISIM '04*, pages 93–100.

R. Valenza, T. Robinson, M. Hickey, and R. Tucker. 1999. Summarization of spoken audio through information extraction. In *Proc. ESCA Workshop on Accessing Information in Spoken Audio*, pages 111–116.

Pierre Wellner, Mike Flynn, Simon Tucker, and Steve Whittaker. 2005. A meeting browser evaluation test. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 2021–2024, New York, NY, USA. ACM Press.

K. Zechner and A. Waibel. 2000. Minimizing word error rate in textual summaries of spoken language. In *Proc. NAACL-2000*.

Evaluating Summaries and Answers: Two Sides of the Same Coin?

Jimmy Lin[1,3] **and Dina Demner-Fushman**[2,3]
[1]College of Information Studies
[2]Department of Computer Science
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
jimmylin@umd.edu, demner@cs.umd.edu

Abstract

This paper discusses the convergence between question answering and multi-document summarization, pointing out implications and opportunities for knowledge transfer in both directions. As a case study in one direction, we discuss the recent development of an automatic method for evaluating definition questions based on *n*-gram overlap, a commonly-used technique in summarization evaluation. In the other direction, the move towards topic-oriented summaries requires an understanding of relevance and topicality, issues which have received attention in the question answering literature. It is our opinion that question answering and multi-document summarization represent two complementary approaches to the same problem of satisfying complex user information needs. Although this points to many exciting opportunities for system-building, here we primarily focus on implications for system evaluation.

1 Introduction

Recent developments in question answering (QA) and multi-document summarization point to many interesting convergences that present exciting opportunities for collaboration and cross-fertilization between these largely independent communities. This position paper attempts to draw connections between the task of answering complex natural language questions and the task of summarizing multiple documents, the boundaries between which are beginning to blur, as anticipated half a decade ago (Carbonell et al., 2000).

Although the complementary co-evolution of question answering and document summarization presents new directions for system-building, this paper primarily focuses on implications for evaluation. Although assessment of answer and summary quality employs different methodologies, there are many lessons that each community can learn from the other. The summarization community has extensive experience in intrinsic metrics based on *n*-gram overlap for automatically scoring system outputs against human-generated reference texts—these techniques would help streamline aspects of question answering evaluation. In the other direction, because question answering has its roots in information retrieval, much work has focused on extrinsic metrics based on relevance and topicality, which may be valuable to summarization researchers.

This paper is organized as follows: In Section 2, we discuss the evolution of question answering research and how recent trends point to the convergence of question answering and multi-document summarization. In Section 3, we present a case study of automatically evaluating definition questions by employing metrics based on *n*-gram overlap, a general technique widely used in summarization and machine translation evaluations. Section 4 highlights some opportunities for knowledge transfer in the other direction: how the notions of rele-

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 41–48, Ann Arbor, June 2005.

vance and topicality, well-studied in the information retrieval literature, can guide the evaluation of topic-oriented summaries. We conclude with thoughts about the future in Section 5.

2 Convergence of QA and Summarization

Question answering was initially conceived as essentially a fine-grained information retrieval task. Much research has focused on so-called factoid questions, which can typically be answered by named entities such as people, organizations, locations, etc. As an example, a system might return "Bee Gees" as the answer to the question "What band did the music for the 1970's film 'Saturday Night Fever'?". For such well-specified information needs, question answering systems represent an improvement over traditional document retrieval systems because they do not require a user to manually browse through a ranked list of "hits". Since 1999, the NIST-organized question answering tracks at TREC (see, for example, Voorhees 2003a) have served as a focal point of research in the field, providing an annual forum for evaluating systems developed by teams from all over the world. The model has been duplicated and elaborated on by CLEF in Europe and NTCIR in Asia, both of which have also introduced cross-lingual elements.

Recently, research in question answering has shifted away from factoid questions to more complex information needs. This new direction can be characterized as a move towards answers that can only be arrived at through some form of reasoning and answers that require drawing information from multiple sources. Indeed, there are many types of questions that would require integration of both capabilities: extracting raw information "nuggets" from potentially relevant documents, reasoning over these basic facts to draw additional inferences, and synthesizing an appropriate answer based on this knowledge. "What is the role of the Libyan government in the Lockerbie bombing?" is an example of such a complex question.

Commonalities between the task of answering complex questions and summarizing multiple documents are evident when one considers broader research trends. Both tasks require the ability to draw together elements from multiple sources and cope with redundant, inconsistent, and contradictory information. Both tasks require extracting finer-grained (i.e., sub-document) segments, albeit based on different criteria. These observations point to the convergence of question answering and multi-document summarization.

Complementary developments in the summarization community mirror the aforementioned shifts in question answering research. Most notably, the DUC 2005 task requires systems to generate answers to natural language questions based on a collection of known relevant documents: "The system task in 2005 will be to synthesize from a set of 25–50 documents a brief, well-organized, fluent answer to a need for information that cannot be met by just stating a name, date, quantity, etc." (DUC 2005 guidelines). These guidelines were modeled after the *information synthesis* task suggested by Amigó et al. (2004), which they characterize as "the process of (given a complex information need) extracting, organizing, and inter-relating the pieces of information contained in a set of relevant documents, in order to obtain a comprehensive, non-redundant report that satisfies the information need". One of the examples they provide, "I'm looking for information concerning the history of text compression both before and with computers", looks remarkably like a user information need current question answering systems aspire to satisfy. The idea of topic-oriented multi-document summarization isn't new (Goldstein et al., 2000), but only recently have the connections to question answering become explicit. Incidentally, it appears that the current vision of question answering is more ambitious than the information synthesis task because in the former, the set of relevant documents is not known in advance, but must first be discovered within a larger corpus.

There is, however, an important difference between question answering and topic-focused multi-document summarization: whereas summaries are compressible in length, the same cannot be said of answers.[1] For question answering, it is difficult to fix the length of a response *a priori*: there may be cases where it is impossible to fit a coherent, complete answer into an allotted space. On the other

[1] We would like to thank an anonymous reviewer for pointing this out.

1	*vital*	american composer
2	*vital*	musical achievements ballets symphonies
3	*vital*	born brooklyn ny 1900
4	*okay*	son jewish immigrant
5	*okay*	american communist
6	*okay*	civil rights advocate
7	*okay*	had senile dementia
8	*vital*	established home for composers
9	*okay*	won oscar for "the Heiress"
10	*okay*	homosexual
11	*okay*	teacher tanglewood music center boston symphony

Table 1: The "answer key" to the question "Who is Aaron Copland?"

hand, summaries are condensed representations of content, and should theoretically be expandable and compressible based on the level of detail desired.

What are the implications, for system evaluations, of this convergence between question answering and multi-document summarization? We believe that the two fields have much to benefit from each other. In one direction, the question answering community currently lacks experience in automatically evaluating unstructured answers, which has been the focus of much research in document summarization. In the other direction, the question answering community, due to its roots in information retrieval, has a good grasp on the notions of relevance and topicality, which are critical to the assessment of topic-oriented summaries. In the next section, we present a case study in leveraging summarization evaluation techniques to automatically evaluate definition questions. Following that, we discuss how lessons from question answering (and more broadly, information retrieval) can be applied to assist in evaluating summarization systems.

3 Definition Questions: A Case Study

Definition questions represent complex information needs that involve integrating facts from multiple documents. A typical definition question is "What is the Cassini space probe?", to which a system might respond with answers that include "interplanetary probe to Saturn", "carries the Huygens probe to study the atmosphere of Titan, Saturn's largest moon", and "a joint project between NASA, ESA, and ASI". The goal of the task is to return as many interesting "nuggets" of information as possible about the target entity being defined (the Cassini space probe, in this case) while minimizing the amount of irrelevant information retrieved. In the two formal evaluations of definition questions that have been conducted at TREC (in 2003 and 2004), an information nugget is operationalized as a fact for which an assessor could make a binary decision as to whether a response contained that nugget (Voorhees, 2003b). Additionally, information nuggets are classified as either *vital* or *okay*. Vital nuggets represent facts central to the target entity, and should be present in a "good" definition. Okay nuggets contribute worthwhile information about the target, but are not essential. As an example, assessors' nuggets for the question "Who is Aaron Copland?" are shown in Table 1. The distinction between vital and okay nuggets is consequential for the score calculation, which we will discuss below.

In the TREC setup, a system response to a definition question is comprised of an unordered set of answer strings paired with the identifier of the document from which it was extracted. Each of these answer strings is presumed to have one or more information nuggets contained within it. Although there is no explicit limit on the length of each answer string and the number of answer strings a system is allowed to return, verbosity is penalized against, as we shall see below.

To evaluate system output, NIST gathers answer strings from all participants, hides their association

[NYT19990708.0196] Once past a rather routine apprenticeship, which included three years of study with Nadia Boulanger in Paris, Copland became one of the few American composers to make a living from composition.
Nugget present: 1

[NYT20000107.0305] A passionate advocate of civil rights, Copland conducted a performance of the "Lincoln Portrait" with Coretta Scott King as narrator.
Nuggets present: 6

[NYT19991117.0369] after four prior nominations, he won an Oscar in 1949 for his music for "The Heiress"
Nugget present: 9

Figure 1: Examples of judging actual system responses.

with the runs that produced them, and presents all answer strings to a human assessor. Using these responses and research performed during the original development of the question (with an off-the-shelf document retrieval system), the assessor creates an "answer key"; Table 1 shows the official answer key for the question "Who is Aaron Copland?".

After this answer key has been created, NIST assessors then go back over each run and manually judge whether or not each nugget is present in a particular system's response. Figure 1 shows a few examples of real system output and the nuggets that were found in them.

The final score of a particular answer is computed as an F-measure, the harmonic mean between nugget precision and recall. The β parameter controls the relative importance of precision and recall, and is heavily biased towards the latter to model the nature of the task. Nugget recall is calculated solely as a function of the vital nuggets, which means that a system receives no "credit" (in terms of recall) for returning okay nuggets. Nugget precision is approximated by a length allowance based on the number of vital and okay nuggets returned; a response longer than the allowed length is subjected to a verbosity penalty. Using answer length as a proxy to precision appears to be a reasonable compromise because a pilot study demonstrated that it was impossible for humans to consistently enumerate the total number of nuggets in a response, a necessary step in calculating nugget precision (Voorhees, 2003b).

The current TREC setup for evaluating definition

Let

r	# of *vital* nuggets returned in a response
a	# of *okay* nuggets returned in a response
R	# of *vital* nuggets in the answer key
l	# of non-whitespace characters in the entire answer string

Then

$$\text{recall } (\mathcal{R}) = r/R$$

$$\text{allowance } (\alpha) = 100 \times (r + a)$$

$$\text{precision } (\mathcal{P}) = \begin{cases} 1 & \text{if } l < \alpha \\ 1 - \frac{l-\alpha}{l} & \text{otherwise} \end{cases}$$

Finally, $F(\beta) = \dfrac{(\beta^2 + 1) \times \mathcal{P} \times \mathcal{R}}{\beta^2 \times \mathcal{P} + \mathcal{R}}$

$\beta = 5$ in TREC 2003, $\beta = 3$ in TREC 2004.

Figure 2: Official definition of F-measure.

questions necessitates having a human "in the loop". Even though answer keys are available for questions from previous years, determining if a nugget was actually retrieved by a system currently requires human judgment. Without a fully-automated evaluation method, it is difficult to consistently and reproducibly assess the performance of a system outside the annual TREC cycle. Thus, researchers cannot carry out controlled laboratory experiments to rapidly explore the solution space. In many other fields in computational linguistics, the ability to conduct evaluations with quick turnaround has lead to rapid progress in the state of the art. Question an-

swering for definition questions appears to be missing this critical ingredient.

To address this evaluation gap, we have recently developed POURPRE, a method for automatically evaluating definition questions based on *idf*-weighted unigram co-occurrences (Lin and Demner-Fushman, 2005). This idea of employing *n*-gram co-occurrence statistics to score the output of a computer system against one or more desired reference outputs has its roots in the BLEU metric for machine translation (Papineni et al., 2002) and the ROUGE (Lin and Hovy, 2003) metric for summarization. Note that metrics for automatically evaluating definitions should be, like metrics for evaluating summaries, biased towards recall. Fluency (i.e., precision) is not usually of concern because most systems employ extractive techniques to produce answers. Our study reports good correlation between the automatically computed POURPRE metric and official TREC system ranks. This measure will hopefully spur progress in definition question answering systems.

The development of automatic evaluation metrics based on *n*-gram co-occurrence for question answering is an example of successful knowledge transfer from summarization to question answering evaluation. We believe that there exist many more opportunities for future exploration; as an example, there are remarkable similarities between information nuggets in definition question answering and recently-proposed methods for assessing summaries based on fine-grained semantic units (Teufel and van Halteren, 2004; Nenkova and Passonneau, 2004).

Another promising direction of research in definition question answering involves applying the Pyramid Method (Nenkova and Passonneau, 2004) to better model the vital/okay nuggets distinction. As it currently stands, the vital/okay dichotomy is troublesome because there is no way to operationalize such a classification scheme within a system; see Hildebrandt et al. (2004) for more discussion. Yet, the effects on score are significant: a system that returns, for example, all the okay nuggets but none of the vital nuggets would receive a score of zero. In truth, the vital/okay distinction is a poor attempt at modeling the fact that some nuggets about a target are more important than others—this is exactly what the Pyramid Method is designed to capture. "Building pyramids" for definition questions is an avenue of research that we are currently pursuing.

In the next section, we discuss opportunities for knowledge transfer in the other direction; i.e., how summarization evaluation can benefit from work in question answering evaluation.

4 Putting the Relevance in Summarization

The definition of a meaningful extrinsic evaluation metric (e.g., a task-based measure) is an issue that the summarization community has long grappled with (Mani et al., 2002). This issue has been one of the driving factors towards summaries that are specifically responsive to complex information needs. The evaluation of such summaries hinges on the notions of relevance and topicality, two themes that have received much research attention in the information retrieval community, from which question answering evolved.

Debates about the nature of relevance are almost as old as the field of information retrieval itself (Cooper, 1971; Saracevic, 1975; Harter, 1992; Barry and Schamber, 1998; Mizzaro, 1998; Spink and Greisdorf, 2001). Theoretical discussions aside, there is evidence suggesting that there exist substantial inter-assessor differences in document-level relevance judgments (Voorhees, 2000; Voorhees, 2002); in the TREC *ad hoc* tracks, for example, overlap between two humans can be less than 50%. For factoid question answering, it has also been shown that the notion of answer correctness is less well-defined than one would expect (Voorhees and Tice, 2000; Lin and Katz, 2005 in press). This inescapable fact about the nature of information needs represents a fundamental philosophical difference between research in information retrieval and computational linguistics. Information retrieval researchers accept the fact that the notion of "ground truth" is not particularly meaningful, and any prescriptive attempt to dictate otherwise would result in brittle and overtrained systems of limited value. A retrieval system must be sensitive to the inevitable variations in relevance exhibited by different users.

This philosophy represents a contrast from computational linguistics research, where ground truth does in fact exist. For example, there is a single correct parse of a natural language sentence (modulo

truly ambiguous sentences), there is the notion of a correct word sense (modulo granularity issues), etc. This view also pervades evaluation in machine translation and document summarization, and is implicitly codified in intrinsic metrics, except that there is now the notion of multiple correct answers (i.e., the reference texts).

Faced with the inevitability of variations in humans' notion of relevance, how can information retrieval researchers confidently draw conclusions about system performance and the effectiveness of various techniques? Meta-evaluations have shown that while some measures such as recall are relatively meaningless in absolute terms (e.g., the total number of relevant documents cannot be known without exhaustive assessment of the entire corpus, which is impractical for current document collections), relative comparisons between systems are remarkably stable. That is, if system A performs better than system B (by a metric such as mean average precision, for example), system A is highly likely to out-perform system B with any alternative sets of relevance judgments that represent different notions of relevance (Voorhees, 2000; Voorhees, 2002). Thus, it remains possible to determine the relative effectiveness of different retrieval techniques, and use evaluation results to guide system development.

We believe that this philosophical starting point for conducting evaluations is an important point that summarization researchers should take to heart, considering that notions such as relevance and topicality are central to the evaluation of the information synthesis task. What concrete implications of this view are there? We outline some thoughts below:

First, we believe that summarization metrics should embrace variations in human judgment as an inescapable part of the evaluation process. Measures for automatically assessing the quality of a system's output such as ROUGE implicitly assume that the "best summary" is a statistical agglomeration of the reference summaries, which is not likely to be true. Until recently, ROUGE "hard-coded" the so-called "jackknifing" procedure to estimate average human performance. Fortunately, it appears researchers have realized that "model averaging" may not be the best way to capture the existence of many "equally good" summaries. As an example, the Pyramid Method (Nenkova and Passonneau, 2004), represents a good first attempt at a realistic model of human variations.

Second, the view that variations in judgment are an inescapable part of extrinsic evaluations would lead one to conclude that low inter-annotator agreement isn't necessarily bad. Computational linguistics research generally attaches great value to high kappa measures (Carletta, 1996), which indicate high human agreement on a particular task. Low agreement is seen as a barrier to conducting reproducible research and to drawing generalizable conclusions. However, this is not necessarily true—low agreement in information retrieval has not been a handicap for advancing the state of the art. When dealing with notions such as relevance, low kappa values can most likely be attributed to the nature of the task itself. Attempting to raise agreement by, for example, developing rigid assessment guidelines, may do more harm than good. Prescriptive attempts to define what a good answer or summary should be will lead to systems that are not useful in real-world settings. Instead, we should focus research on adaptable, flexible systems.

Third, meta-evaluations are important. The information retrieval literature has an established tradition of evaluating evaluations post hoc to insure the reliability and fairness of the results. The aforementioned studies examining the impact of different relevance judgments are examples of such work. Due to the variability in human judgments, systems are essentially aiming at a moving target, which necessitates continual examination as to whether evaluations are accurately answering the research questions and producing trustworthy results.

Fourth, a measure for assessing the quality of automatic scoring metrics should reflect the philosophical starting points that we have been discussing. As a specific example, the correlation between an automatically-calculated metric and actual human preferences is better quantified by Kendall's τ than by the coefficient of determination R^2. Since relative system comparisons are more meaningful than absolute scores, we are generally less interested in correlations among the scores than in the rankings of systems produced by those scores. Kendall's τ computes the "distance" between two rankings as the minimum number of pairwise adjacent swaps necessary to convert one ranking into the other. This value

is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0; the correlation between a ranking and its perfect inverse is -1.0; and the expected correlation of two rankings chosen at random is 0.0. Typically, a value of greater than 0.8 is considered "good", although 0.9 represents a threshold researchers generally aim for.

5 Conclusion

What's in store for the ongoing co-evolution of summarization and question answering? Currently, definition questions exercise a system's ability to integrate information from multiple documents. In the process, it needs to automatically recognize similar information units to avoid redundant information, much like in multi-document summarization. The other research direction in advanced question answering, integration of reasoning capabilities to generate answers that cannot be directly extracted from text, remains more elusive for a variety of reasons. Finer-grained linguistic analysis at a large scale and sufficiently-rich domain ontologies to support potentially long inference chains are necessary prerequisites—both of which represent open research problems. Furthermore, it is unclear how exactly one would operationalize the evaluation of such capabilities.

Nevertheless, we believe that advanced reasoning capabilities based on detailed semantic analyses of text will receive much attention in the future. The recent flurry of work on semantic analysis, based on resources such as FrameNet (Baker et al., 1998) and PropBank (Kingsbury et al., 2002), provide the substrate for reasoning engines. Developments in the automatic construction, adaptation, and merging of ontologies will supply the knowledge necessary to draw inferences. In order to jump-start the knowledge acquisition process, we envision the development of domain-specific question answering systems, the lessons from which will be applied to systems that operate on broader domains. In terms of operationalizing evaluations for these advanced capabilities, the field has already made important first steps, e.g., the Pascal Recognising Textual Entailment Challenge.

What effect will these developments have on summarization research? We believe that future systems will employ more detailed linguistic analysis. As a simple example, the ability to reason about people's age based on their birthdates would undoubtedly be useful for answering particular types of questions, but may also play a role in redundancy detection, for example. In general, we anticipate a move towards more abstractive techniques in multi-document summarization. Fluent, cohesive, and topical summaries cannot be generated solely using an extractive approach—sentences are at the wrong level of granularity, a source of problems ranging from dangling anaphoric references to verbose subordinate clauses. Only through more detailed linguistic analysis can information from multiple documents be truly synthesized. Already, there are hybrid approaches to multi-document summarization that employ natural language generation techniques (McKeown et al., 1999; Elson, 2004), and researchers have experimented with sentential operations to improve the discourse structure of summaries (Otterbacher et al., 2002).

The primary purpose of this paper was to identify similarities between multi-document summarization and complex question answering, pointing out potential synergistic opportunities in the area of system evaluation. We hope that this is merely a small part of a sustained dialogue between researchers from these two largely independent communities. Answering complex questions and summarizing multiple documents are essentially opposite sides of the same coin, as they represent different approaches to the common problem of addressing complex user information needs.

6 Acknowledgements

We would like to thank Donna Harman and Ellen Voorhees for many insights about the intricacies of IR evaluation, Bonnie Dorr for introducing us to DUC and bringing us into the summarization community, and Kiri for her kind support.

References

Enrique Amigó, Julio Gonzalo, Victor Peinado, Anselmo Peñas, and Felisa Verdejo. 2004. An empirical study of information synthesis task. In *Proceedings of ACL 2004*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING/ACL 1998*.

Carol Barry and Linda Schamber. 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34(2/3):219–236.

Jaime Carbonell, Donna Harman, Eduard Hovy, Steve Maiorano, John Prange, and Karen Sparck-Jones. 2000. Vision statement to guide research in Question & Answering (Q&A) and Text Summarization.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

William S. Cooper. 1971. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37.

David K. Elson. 2004. Categorization of narrative semantics for use in generative multidocument summarization. In *Proceedings of INLG 2004*, pages 192–197.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Jamie Callan. 2000. Creating and evaluating multidocument sentence extract summaries. In *Proceedings of CIKM 2000*.

Stephen P. Harter. 1992. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9):602–615.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of HLT/NAACL 2004*.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceeding of HLT 2002*.

Jimmy Lin and Dina Demner-Fushman. 2005. Automatically evaluating answers to definition questions. Technical Report LAMP-TR-119/CS-TR-4695/UMIACS-TR-2005-04, University of Maryland, College Park.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT/NAACL 2003*.

Jimmy Lin and Boris Katz. 2005, in press. Building a reusable test collection for question answering. *Journal of the American Society for Information Science and Technology*.

Inderjeet Mani, Therese Firmin, David House, Gary Klein, Beth Sundheim, and Lynette Hirschman. 2002. The TIPSTER SUMMAC text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Kathleen R. McKeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of AAAI-1999*.

Stefano Mizzaro. 1998. How many relevances in information retrieval? *Interacting With Computers*, 10(3):305–322.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The Pyramid Method. In *Proceedings of HLT/NAACL 2004*.

Jahna C. Otterbacher, Dragomir R. Radev, and Airong Luo. 2002. Revisions that improve cohesion in multidocument summaries: A preliminary study. In *Proceedings of the ACL 2002 Workshop on Automatic Summarization*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Tefko Saracevic. 1975. Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343.

Amanda H. Spink and Howard Greisdorf. 2001. Regions and levels: Mapping and measuring users relevance judgments. *Journal of the American Society for Information Science and Technology*, 52(2):161–173.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of EMNLP 2004*.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of SIGIR 2000*.

Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716.

Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems, Springer-Verlag LNCS 2406*.

Ellen M. Voorhees. 2003a. Evaluating the evaluation: A case study using the TREC 2002 question answering track. In *Proceedings of HLT/NAACL 2003*.

Ellen M. Voorhees. 2003b. Overview of the TREC 2003 question answering track. In *Proceedings of TREC 2003*.

Evaluating DUC 2004 Tasks with the QARLA Framework

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
c/Juan del Rosal, 16 - 28040 Madrid - Spain
{enrique,julio,anselmo,felisa}@lsi.uned.es

Abstract

This papers reports the application of the QARLA evaluation framework to the DUC 2004 testbed (tasks 2 and 5). Our experiment addresses two issues: how well QARLA evaluation measures correlate with human judgements, and what additional insights can be provided by the QARLA framework to the DUC evaluation exercises.

1 Introduction

QARLA (Amigó et al., 2005) is a framework that uses similarity to models as a building block for the evaluation of automatic summarisation systems. The input of QARLA is a summarisation task, a set of test cases, a set of similarity metrics, and sets of models and automatic summaries (peers) for each test case. With such a testbed, QARLA provides:

- A measure, QUEEN, which combines assorted similarity metrics to estimate the quality of automatic summarisers.
- A measure, KING, to select the best combination of similarity metrics.
- An estimation, JACK, of the reliability of the testbed for evaluation purposes.

The QARLA framework does not rely on human judges. It is interesting, however, to find out how well an evaluation using QARLA correlates with human judges, and whether QARLA can provide additional insights into an evaluation based on human assessments.

In this paper, we apply the QARLA framework (QUEEN, KING and JACK measures) to the output of two different evaluation exercises: DUC 2004 tasks 2 and 5 (Over and Yen, 2004). Task 2 requires short (one-hundred word) summaries for assorted document sets; Task 5 consists of generating a short summary in response to a "Who is" question.

In Section 2, we summarise the QARLA evaluation framework; in Section 3, we describe the similarity metrics used in the experiments. Section 4 discusses the results of the QARLA framework using such metrics on the DUC testbeds. Finally, Section 5 draws some conclusions.

2 The QARLA evaluation framework

QARLA uses similarity to models for the evaluation of automatic summarisation systems. Here we summarise its main features; the reader may refer to (Amigó et al., 2005) for details.

The input of the framework is:

- A summarisation task (e.g. topic oriented, informative multi-document summarisation on a given domain/corpus).
- A set T of test cases (e.g. topic/document set pairs for the example above)
- A set of summaries M produced by humans (*models*), and a set of automatic summaries A (*peers*), for every test case.
- A set X of similarity metrics to compare summaries.

With this input, QARLA provides three main measures that we describe below.

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 49–56, Ann Arbor, June 2005.

2.1 $QUEEN$: Estimating the quality of an automatic summary

QUEEN operates under the assumption that a summary is better if it is closer to the model summaries according to all metrics; it is defined as the probability, measured on $M \times M \times M$, that for every metric in X the automatic summary a is closer to a model than two models to each other:

$$\text{QUEEN}_{X,M}(a) \equiv P(\forall x \in X . x(a,m) \geq x(m',m''))$$

where a is the automatic summary being evaluated, $\langle m, m', m'' \rangle$ are three models in M, and $x(a,m)$ stands for the similarity of m to a. QUEEN is stated as a probability, and therefore its range of values is $[0,1]$.

We can think of the QUEEN measure as using a set of tests (every similarity metric in X) to falsify the hypothesis that a given summary a is a model. Given $\langle a, m, m', m'' \rangle$, we test $x(a,m) \geq x(m',m'')$ for each metric x. a is accepted as a model only if it passes the test for every metric. QUEEN(a) is, then, the probability of acceptance for a in the sample space $M \times M \times M$.

This measure has some interesting properties: **(i)** it is able to combine different similarity metrics into a single evaluation measure; **(ii)** it is not affected by the scale properties of individual metrics, i.e. it does not require metric normalisation and it is not affected by metric weighting. **(iii)** Peers which are very far from the set of models all receive QUEEN=0. In other words, QUEEN does not distinguish between very poor summarisation strategies. **(iv)** The value of QUEEN is maximised for peers that "merge" with the models under all metrics in X. **(v)** The universal quantifier on the metric parameter x implies that adding redundant metrics do not bias the result of QUEEN.

Now the question is: which similarity metrics are adequate to evaluate summaries? Imagine that we use a similarity metric based on sentence co-selection; it might happen that humans do not agree on which sentences to select, and therefore emulating their sentence selection behaviour is both easy (nobody agrees with each other) and useless. We need to take into account which are the features that human summaries do share, and evaluate according to them. This is provided by the KING measure.

2.2 $KING$: estimating the quality of similarity metrics

The measure KING$_{M,A}(X)$ estimates the quality of a set of similarity metrics X using a set of models M and a set of peers A. KING is defined as the probability that a model has higher QUEEN value than any peer in a test sample. Formally:

$$\text{KING}_{M,A}(X) \equiv$$

$$P(\forall a \in A, \text{QUEEN}_{M,X}(m) > \text{QUEEN}_{M,X}(a))$$

For example, an ideal metric -that puts all models together-would give QUEEN$(m) = 1$ for all models, and QUEEN$(a) = 0$ for all peers which are not put together with the models, obtaining KING $= 1$.

KING satisfies several interesting properties: (i) KING does not depend on the scale properties of the metric; (ii) Adding repeated or very similar peers do not alter the KING measure, which avoids one way of biasing the measure. (iii) the KING value of random and constant metrics is zero or close to zero.

2.3 JACK: reliability of the peer set

Once we detect a difference in quality between two summarisation systems, the question is now whether this result is reliable. Would we get the same results using a different test set (different examples, different human summarisers (models) or different baseline systems)?

The first step is obviously to apply statistical significance tests to the results. But even if they give a positive result, it might be insufficient. The problem is that the estimation of the probabilities in KING assumes that the sample sets M, A are not biased. If M, A are biased, the results can be statistically significant and yet unreliable. The set of examples and the behaviour of human summarisers (models) should be somehow controlled either for homogeneity (if the intended profile of examples and/or users is narrow) or representativity (if it is wide). But how to know whether the set of automatic summaries is representative and therefore is not penalising certain automatic summarisation strategies?

This is addressed by the JACK measure:

$$\text{JACK}(X, M, A) \equiv P(\exists a, a' \in A|$$
$$\forall x \in X. x(a, a') \leq x(a, m) \wedge x(a', a) \leq x(a', m) \wedge$$
$$\text{QUEEN}(a) > 0 \wedge \text{QUEEN}(a') > 0)$$

i.e. the probability over all model summaries m of finding a couple of automatic summaries a, a' which are closer to m than to each other according to all metrics. This measure satisfies three desirable properties: (i) it can be enlarged by increasing the similarity of the peers to the models (the $x(m, a)$ factor in the inequalities), i.e. enhancing the quality of the peer set; (ii) it can also be enlarged by decreasing the similarity between automatic summaries (the $x(a, a')$ factor in the inequality), i.e. augmenting the diversity of (independent) automatic summarisation strategies represented in the test bed; (iii) adding elements to A cannot diminish the JACK value, because of the existential quantifier on a, a'.

3 Selection of similarity metrics

Each different similarity metric characterises different features of a summary. Our first objective is to select the best set of metrics, that is, the metrics which best characterise the human summaries (models) as opposed to automatic summaries. The second objective is to obtain as much information as possible about the behaviour of automatic summaries.

In this Section, we begin by describing a set of 59 metrics used as a starting point. Some of them provide overlapping information; the second step is then to select a subset of metrics that minimises redundancy and, at the same time, maximises quality (KING values). Finally, we analyse the characteristics of the selected metrics.

3.1 Similarity metrics

For this work, we have considered the following similarity metrics:

ROUGE based metrics (R): ROUGE (Lin and Hovy, 2003) estimates the quality of an automatic summary on the basis of the n-gram coverage related to a set of human summaries (models). Although ROUGE is an evaluation metric, we can adapt it to behave as a similarity metric between pairs of summaries if we consider only one model in the computation. There are different kinds of ROUGE metrics such as ROUGE-W, ROUGE-L, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, etc. (Lin, 2004b). Each of these metrics has been applied over summaries with three preprocessing options: with stemming and stopword removal (type c); only with stopwords removal (type b); or without any kind of preprocessing (type a). All these combinations give 24 similarity metrics based on ROUGE.

Inverted ROUGE based metrics (Rpre): ROUGE metrics are recall oriented. If we reverse the direction of the similarity computation, we obtain precision oriented metrics (i.e. $\text{Rpre}(a, b) = \text{R}(b, a)$). In this way, we generate another 24 metrics based on inverted ROUGE.

TruncatedVectModel (TVM$_n$): This family of metrics compares the distribution of the n most relevant terms from original documents in the summaries. The process is the following: (1) obtaining the n most frequent lemmas ignoring stopwords; (2) generating a vector with the relative frequency of each term in the summary; (3) calculating the similarity between two vectors as the inverse of the Euclidean distance. We have used 9 variants of this measure with $n = 1, 4, 8, 16, 32, 64, 128, 256, 512$.

AveragedSentencelengthSim (AVLS): This is a very simple metric that compares the average length of the sentences in two summaries. It can be useful to compare the degree of abstraction of the summaries.

GRAMSIM: This similarity metric compares the distribution of the part-of-speech tags in the two summaries. The processing is the following: (1) part-of-speech tagging of summaries using TreeTagger ; (2) generation of a vector with the tags frequency for each summary; (3) calculation of the similarity between two vectors as the inverse of the Euclidean distance. This similarity metric is not content oriented, but syntax-oriented.

cluster ID	DESCRIPTION	SIMILARITY METRICS
Cluster 1	ROUGE based metrics	R-S.b R-SU.b R-S.a R-SU.a R-1.a R-1.b R-L.b R-L.a R-W-1.2.b R-W-1.2.a R-W-1.2.c R-S.c R-SU.c R-1.c R-L.c **Rpre-W-1.2.b** Rpre-W-1.2.a Rpre-W-1.2.c Rpre-L.c Rpre-1.c Rpre-S.c Rpre-SU.c Rpre-1.a Rpre-S.a Rpre-SU.a Rpre-1.b Rpre-S.b Rpre-SU.b Rpre-L.b Rpre-L.a
Cluster 2	ROUGE (Stemmed and non-stopwords 2-grams)	**R-2.c** Rpre-2.c
Cluster 3	ROUGE (Stemmed and non-stopwords 3-grams)	Rpre-3.c **R-3.c**
Cluster 4	ROUGE (Stemmed and non-stopwords 4-grams)	Rpre-4.c **R-4.c**
Cluster 5	ROUGE (Non-stemmed 2-grams)	R-2.b R-2.a **Rpre-2.b** Rpre-2.a
Cluster 6	ROUGE (Non-stemmed 3-grams)	R-3.b R-3.a **Rpre-3.b** Rpre-3.a
Cluster 7	ROUGE (Non-stemmed 4-grams)	Rpre-4.a **Rpre-4.b** R-4.b R-4.a
Cluster 8	TVM.Most salient term	**TVM.1**
Cluster 9	TVM.4 and 8 salient terms	TVM.4 **TVM.8**
Cluster 10	TVM.>8 Salient terms	TVM.16 TVM.32 TVM.64 TVM.128 TVM.256 **TVM.512**

Figure 1: Similarity Metric Clusters

3.2 Clustering similarity metrics

From the set of metrics described above we have 57 (24+24+9) content oriented metrics, plus two metrics based on stylistic features (AVLS and GRAM-SIM). However, the 57 metrics characterising summary contents are highly redundant. Thus, clustering similar metrics seems desirable.

We perform an automatic clustering process using the following notion of proximity between two metric sets:

$$sim(X, X') \equiv Prob[H(X) \leftrightarrow H(X')]$$

$$\text{where } H(X) \equiv \forall x \in X.x(a, m) \geq x(m', m'')$$

Two metrics sets are similar, according to the formula, if they behave similarly with respect to the $QUEEN$ condition (H predicate in the formula), i.e. the probability that the two sets of metrics discriminate the same automatic summaries when they are compared to the same pair of models.

Figure 1 shows the clustering of similarity metrics for the DUC 2004 Task 2. The number of clusters was fixed in 10. After the clustering process, the 48 ROUGE metrics are grouped in 7 sets, and the 9 TVM metrics are grouped in 3 sets. In each cluster, the metric with highest KING has been marked in boldface. Note that the ROUGE-c metrics (with stemming) with highest KING are those based on recall whereas the ROUGE-a/b metrics (without stemming) are those based on precision. Regarding TVM clusters, the metrics with highest KING in each cluster are those based on a higher number of terms.

Finally, we select the metric with highest KING in each group, obtaining the 10 most representative metrics.

3.3 Best evaluation metric: KING values

Figure 2 shows the KING values for the selected similarity metrics, which represent how every metric characterises model summaries as opposed to automatic summaries. These are the main results:

- The last column shows the best metric set, considering all possible metric combinations. In both DUC tasks, the best combination is {Rpre-W-1.2.b, TVM.512. This metric set gets better KING values than any individual metric in isolation (17% better than the second best for task 2, and 23% better for task 5). This is an interesting result confirming that we can improve our ability to characterise human summaries just by combining standard similarity metrics in the QARLA framework. Note also that both metrics in the best set are content-oriented.

- Rpre-W.1.2.b (inverted ROUGE measure, using non-contiguous word sequences, removing stopwords, without stemming) obtains the highest individual KING for task 2, and is one of the best in task 5, confirming that ROUGE-based metrics are a robust way of evaluating summaries, and indicating that non-contiguous word sequences can be more useful for evaluation purposes than n-grams.

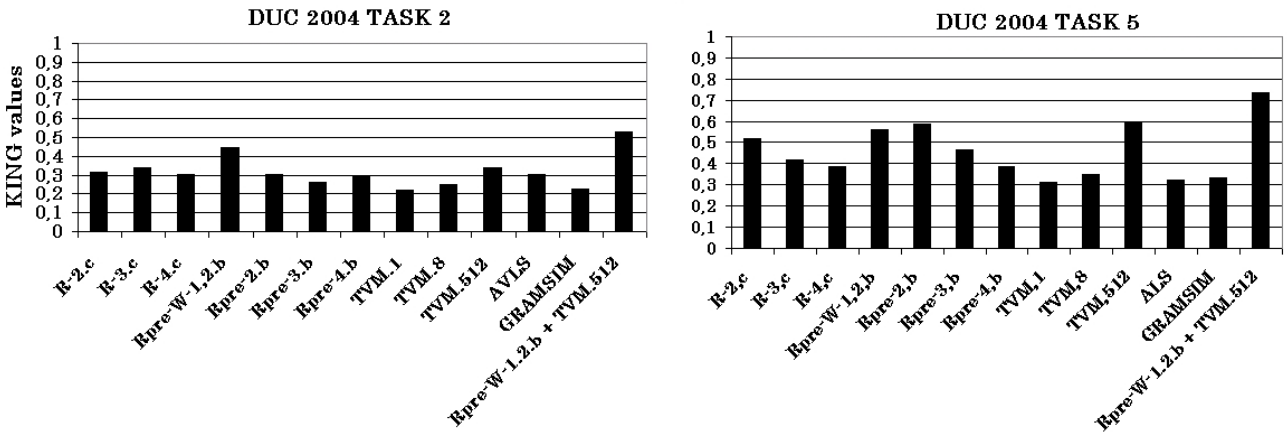


Figure 2: Similarity Metric quality

- TVM metrics get higher values when considering more terms (TVM.512), confirming that comparing with just a few terms (e.g. TVM.4) is not informative enough.

- Overall, KING values are higher for task 5, suggesting that there is more agreement between human summaries in topic-oriented tasks.

3.4 Reliability of the results

The JACK measure estimates the reliability of QARLA results, and is correlated with the diversity of automatic summarisation strategies included in the testbed. In principle, the larger the number of automatic summaries, the higher the JACK values we should obtain. The important point is to determine when JACK values tend to stabilise; at this point, it is not useful to add more automatic summaries without introducing new summarisation strategies.

Figure 3 shows how $JACK_{\text{Rpre-W,TVM.512}}$ values grow when adding automatic summaries. For more than 10 systems, JACK values grow slower in both tasks. Absolute JACK values are higher in Task 2 than in task 5, indicating that systems tend to produce more similar summaries in Task 5 (perhaps because it is a topic-oriented task). This result suggests that we should incorporate more diverse summarisation strategies in Task 5 to enhance the reliability of the testbed for evaluation purposes with QARLA.

4 Evaluation of automatic summarisers: QUEEN values

The QUEEN measure provides two kinds of information to compare automatic summarisation systems: which are the best systems -according to the best metric set-, and which are the individual features of every automatic summariser -according to individual similarity metrics-.

4.1 System ranking

The best metric combination for both tasks was $\{\text{Rpre-W}, \text{TVM.512}\}$; therefore, our global system evaluation uses this combination of content-oriented metrics. Figure 4 shows the $\text{QUEEN}_{\{\text{Rpre-W,TVM.512}\}}$ values for each participating system in DUC 2004, also including the model summaries. As expected, model summaries obtain the highest QUEEN values in both DUC tasks, with a significant distance with respect to the automatic summaries.

4.2 Correlation with human judgements

The manual ranking generated in DUC is based on a set of human-produced evaluation criteria, whereas the QARLA framework gives more weight to the aspects that characterise model summaries as opposed to automatic summaries. It is interesting, however, to find out whether both evaluation methodologies are correlated. Indeed, this is the case: the Pearson correlation between manual and QUEEN rankings is 0.92 for the Task 2 and 0.96 for the Task 5.

Of course, QUEEN values depend on the chosen metric set X; it is also interesting to check whether

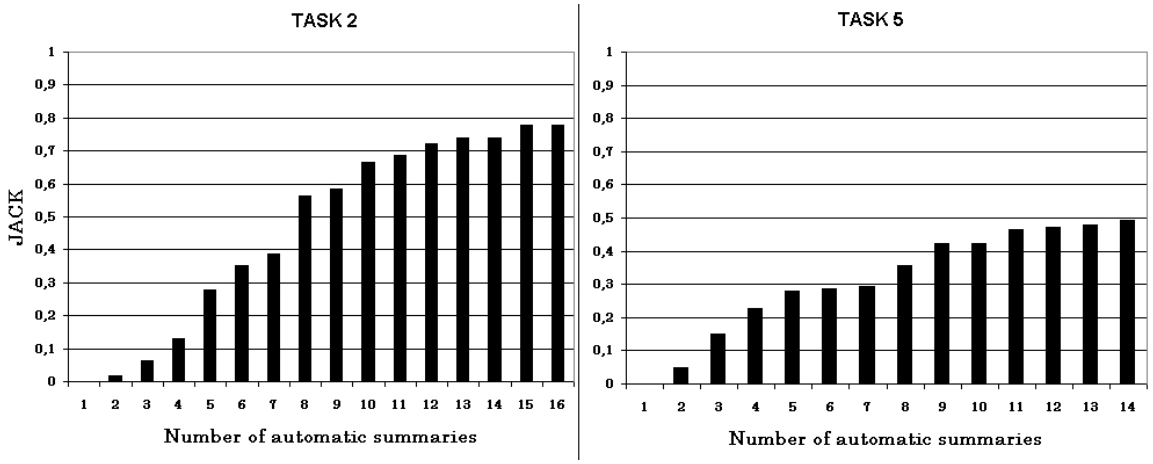


Figure 3: JACK vs. Number of Automatic Summaries

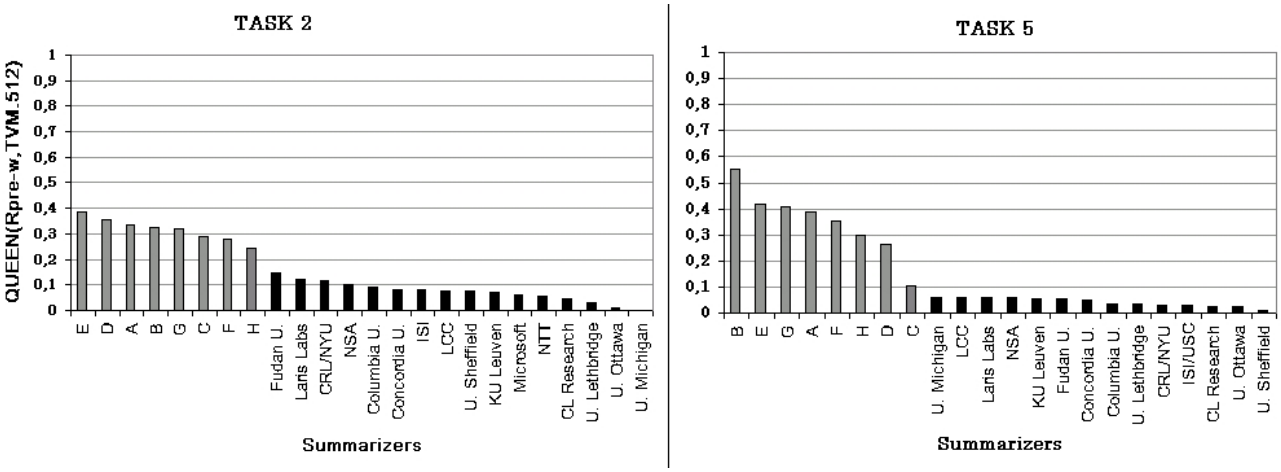


Figure 4: QUEEN system ranking for the best metric set (A-H are models)

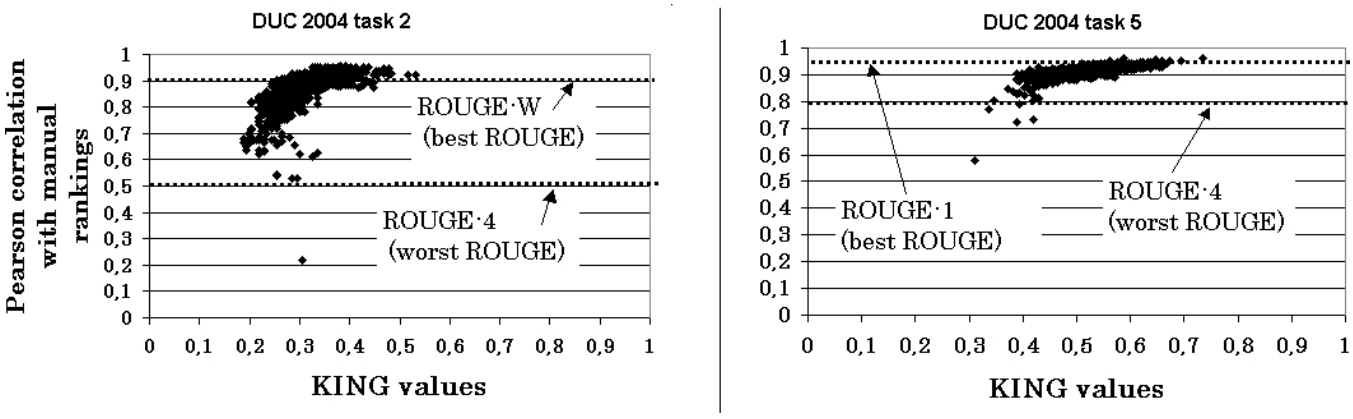


Figure 5: Correlation Between DUC and QARLA results

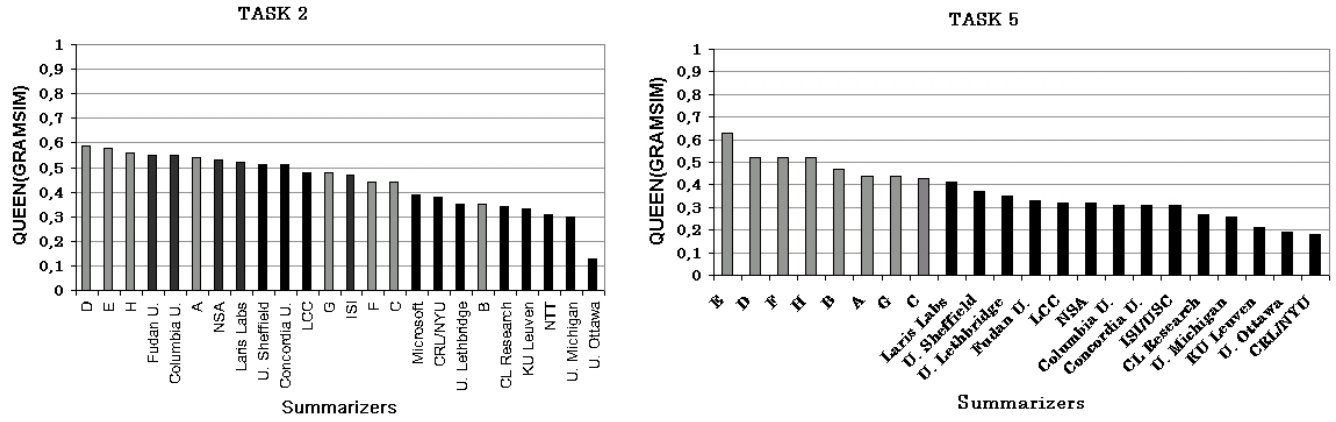


Figure 6: QUEEN values over GRAMSIM

metrics with higher KING values lead to QUEEN rankings more similar to human judgements. Figure 5 shows the Pearson correlation between manual and QUEEN rankings for 1024 metric combinations with different KING values. The figure confirms that higher KING values are associated with rankings closer to human judgements.

4.3 Stylistic features

The best metric combination leaves out similarity metrics based on stylistic features. It is interesting, however, to see how automatic summaries behave with respect to this kind of features. Perhaps the most remarkable fact about stylistic similarities is that, in the case of the GRAMSIM metric, task 2 and task 5 exhibit a rather different behaviour (see Figure 6). In task 2, systems merge with the models, while in task 5 the QUEEN values of the systems are inferior to the models. This suggests that there is some stylistic component in models that systems are not capturing in the topic-oriented task.

5 Related work

The methodology which is closest to our framework is ORANGE (Lin, 2004a), which evaluates a similarity metric using the average ranks obtained by reference items within a baseline set. As in our framework, ORANGE performs an automatic meta-evaluation, there is no need for human assessments, and it does not depend on the scale properties of the metric being evaluated (because changes of scale preserve rankings). The ORANGE approach is, indeed, intimately related to the original QARLA measure introduced in (Amigo et al., 2004).

There are several approaches to the automatic evaluation of summarisation and Machine Translation systems (Culy and Riehemann, 2003; Coughlin, 2003). Probably the most significant improvement over ORANGE is the ability to combine automatically the information of different metrics. Our impression is that a comprehensive automatic evaluation of a summary must necessarily capture different aspects of the problem with different metrics, and that the results of every individual checking (metric) should not be combined in any prescribed algebraic way (such as a linear weighted combination). Our framework satisfies this condition.

ORANGE, however, has also an advantage over the QARLA framework, namely that it can be used for evaluation metrics which are not based on similarity between model/peer pairs. For instance, ROUGE can be applied directly in the ORANGE framework without any reformulation.

6 Conclusions

The application of the QARLA evaluation framework to the DUC testbed provides some useful insights into the problem of evaluating text summarisation systems:

- The results show that a combination of similarity metrics behaves better than any metric in isolation. The best metric set is $\{\text{Rpre-W}, \text{TVM.512}\}$, a combination of content-oriented metrics. Un-

surprisingly, stylistic similarity is less useful for evaluation purposes.

- The evaluation provided by QARLA correlates well with the rankings provided by DUC human judges. For both tasks, metric sets with higher KING values slightly outperforms the best ROUGE evaluation measure.

- QARLA measures show that DUC tasks 2 and 5 are quite different in nature. In Task 5, human summaries are more similar, and the automatic summarisation strategies evaluated are less diverse.

Acknowledgements

We are indebted to Ed Hovy, Donna Harman, Paul Over, Hoa Dang and Chin-Yew Lin for their inspiring and generous feedback at different stages in the development of QARLA. We are also indebted to NIST for hosting Enrique Amigó as a visitor and for providing the DUC test beds. This work has been partially supported by the Spanish government, project R2D2 (TIC-2003-7180).

References

E. Amigó, J. Gonzalo, A. Peñas, and F. Verdejo. 2005. QARLA: a Framework for the Evaluation of Text Summarization Systems. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

E. Amigo, V. Peinado, J. Gonzalo, A. Peñas, and F. Verdejo. 2004. An Empirical Study of Information Synthesis Tasks. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, July.

Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In *In Proceedings of MT Summit IX*, New Orleans,LA.

Christopher Culy and Susanne Riehemann. 2003. The Limits of N-Gram Translation Evaluation Metrics. In *Proceedings of MT Summit IX*, New Orleans,LA.

C. Lin and E. H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-ocurrence Statistics. In *Proceeding of 2003 Language Technology Conference (HLT-NAACL 2003)*.

C. Lin. 2004a. Orange: a Method for Evaluating Automatic Metrics for Machine Translation. In *Proceedings of the 36th Annual Conference on Computational Linguisticsion for Computational Linguistics (Coling'04)*, Geneva, August.

Chin-Yew Lin. 2004b. Rouge: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

P. Over and J. Yen. 2004. An introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems. In *Proceedings of DUC 2004 Document Understanding Workshop, Boston*.

On Some Pitfalls in Automatic Evaluation and Significance Testing for MT

Stefan Riezler and **John T. Maxwell III**
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304

Abstract

We investigate some pitfalls regarding the discriminatory power of MT evaluation metrics and the accuracy of statistical significance tests. In a discriminative reranking experiment for phrase-based SMT we show that the NIST metric is more sensitive than BLEU or F-score despite their incorporation of aspects of fluency or meaning adequacy into MT evaluation. In an experimental comparison of two statistical significance tests we show that p-values are estimated more conservatively by approximate randomization than by bootstrap tests, thus increasing the likelihood of type-I error for the latter. We point out a pitfall of randomly assessing significance in multiple pairwise comparisons, and conclude with a recommendation to combine NIST with approximate randomization, at more stringent rejection levels than is currently standard.

1 Introduction

Rapid and accurate detection of result differences is crucial in system development and system benchmarking. In both situations a multitude of systems or system variants has to be evaluated, so it is highly desirable to employ automatic evaluation measures for detection of result differences, and statistical hypothesis tests to assess the significance of the detected differences. When evaluating subtle differences between system variants in development, or when benchmarking multiple systems, result differences may be very small in magnitude. This imposes strong requirements on both automatic evaluation measures and statistical significance tests: Evaluation measures are needed that have high discriminative power and yet are sensitive to the interesting aspects of the evaluation task. Significance tests are required to be powerful and yet accurate, i.e., if there are significant differences they should be able to assess them, but not if there are none.

In the area of statistical machine translation (SMT), recently a combination of the BLEU evaluation metric (Papineni et al., 2001) and the bootstrap method for statistical significance testing (Efron and Tibshirani, 1993) has become popular (Och, 2003; Kumar and Byrne, 2004; Koehn, 2004b; Zhang et al., 2004). Given the current practice of reporting result differences as small as .3% in BLEU score, assessed at confidence levels as low as 70%, questions arise concerning the sensitivity of the employed evaluation metrics and the accuracy of the employed significance tests, especially when result differences are small. We believe that is important to accurately detect such small-magnitude differences in order to understand how to improve systems and technologies, even though such differences may not matter in current applications.

In this paper we will investigate some pitfalls that arise in automatic evaluation and statistical significance testing in MT research. The first pitfall concerns the discriminatory power of automatic evaluation measures. In the following, we compare the sensitivity of three intrinsic evaluation measures that differ with respect to their focus on different aspects

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 57–64, Ann Arbor, June 2005.

of translation. We consider the well-known BLEU score (Papineni et al., 2001) which emphasizes fluency by incorporating matches of high n-grams. Furthermore, we consider an F-score measure that is adapted from dependency-based parsing (Crouch et al., 2002) and sentence-condensation (Riezler et al., 2003). This measure matches grammatical dependency relations of parses for system output and reference translations, and thus emphasizes semantic aspects of translational adequacy. As a third measure we consider NIST (Doddington, 2002), which favors lexical choice over word order and does not take structural information into account. On an experimental evaluation on a reranking experiment we found that only NIST was sensitive enough to detect small result differences, whereas BLEU and F-score produced result differences that were statistically not significant. A second pitfall addressed in this paper concerns the relation of power and accuracy of significance tests. In situations where the employed evaluation measure produces small result differences, the most powerful significance test is demanded to assess statistical significance of the results. However, accuracy of the assessments of significance is seldom questioned. In the following, we will take a closer look at the bootstrap test and compare it with the related technique of approximate randomization (Noreen (1989)). In an experimental evaluation on our reranking data we found that approximate randomization estimated p-values more conservatively than the bootstrap, thus increasing the likelihood of type-I error for the latter test. Lastly, we point out a common mistake of randomly assessing significance in multiple pairwise comparisons (Cohen, 1995). This is especially relevant in k-fold pairwise comparisons of systems or system variants where k is high. Taking this multiplicity problem into account, we conclude with a recommendation of a combination of NIST for evaluation and the approximate randomization test for significance testing, at more stringent rejection levels than is currently standard in the MT literature. This is especially important in situations where multiple pairwise comparisons are conducted, and small result differences are expected.

2 The Experimental Setup: Discriminative Reranking for Phrase-Based SMT

The experimental setup we employed to compare evaluation measures and significance tests is a discriminative reranking experiment on 1000-best lists of a phrase-based SMT system. Our system is a re-implementation of the phrase-based system described in Koehn (2003), and uses publicly available components for word alignment (Och and Ney, 2003)[1], decoding (Koehn, 2004a)[2], language modeling (Stolcke, 2002)[3] and finite-state processing (Knight and Al-Onaizan, 1999)[4]. Training and test data are taken from the Europarl parallel corpus (Koehn, 2002)[5].

Phrase-extraction follows Och et al. (1999) and was implemented by the authors: First, the word aligner is applied in both translation directions, and the intersection of the alignment matrices is built. Then, the alignment is extended by adding immediately adjacent alignment points and alignment points that align previously unaligned words. From this many-to-many alignment matrix, phrases are extracted according to a contiguity requirement that states that words in the source phrase are aligned only with words in the target phrase, and vice versa.

Discriminative reranking on a 1000-best list of translations of the SMT system uses an ℓ_1 regularized log-linear model that combines a standard maximum-entropy estimator with an efficient, incremental feature selection technique for ℓ_1 regularization (Riezler and Vasserman, 2004). Training data are defined as pairs $\{(s_j, t_j)\}_{j=1}^{m}$ of source sentences s_j and gold-standard translations t_j that are determined as the translations in the 1000-best list that best match a given reference translation. The objective function to be minimized is the conditional log-likelihood $L(\boldsymbol{\lambda})$ subject to a regularization term $R(\boldsymbol{\lambda})$, where $T(s)$ is the set of 1000-best translations for sentence s, $\boldsymbol{\lambda}$ is a vector or log-parameters, and

[1] http://www.fjoch.com/GIZA++.html
[2] http://www.isi.edu/licensed-sw/pharaoh/
[3] http://www.speech.sri.com/projects/srilm/
[4] http://www.isi.edu/licensed-sw/carmel/
[5] http://people.csail.mit.edu/people/koehn/publications/europarl/

Table 1: NIST, BLEU, F-scores for reranker and baseline on development set

	NIST	BLEU	F
baseline	6.43	.301	.385
reranking	6.58	.298	.383
approxrand p-value	$< .0001$	.158	.424
bootstrap p-value	$< .0001$	.1	-

$\boldsymbol{f}$ is a vector of feature functions:

$$\begin{aligned} L(\boldsymbol{\lambda}) + R(\boldsymbol{\lambda}) &= -\log \prod_{j=1}^{m} p_{\boldsymbol{\lambda}}(t_j|s_j) + R(\boldsymbol{\lambda}) \\ &= -\sum_{j=1}^{m} \log \frac{e^{\boldsymbol{\lambda}\cdot\boldsymbol{f}(t_j)}}{\sum_{t\in T(s_j)} e^{\boldsymbol{\lambda}\cdot\boldsymbol{f}(t)}} + R(\boldsymbol{\lambda}) \end{aligned}$$

The features employed in our experiments consist of 8 features corresponding to system components (distortion model, language model, phrase-translations, lexical weights, phrase penalty, word penalty) as provided by PHARAOH, together with a multitude of overlapping phrase features. For example, for a phrase-table of phrases consisting of maximally 3 words, we allow all 3-word phrases and 2-word phrases as features. Since bigram features can overlap, information about trigrams can be gathered by composing bigram features even if the actual trigram is not seen in the training data.

Feature selection makes it possible to employ and evaluate a large number of features, without concerns about redundant or irrelevant features hampering generalization performance. The ℓ_1 regularizer is defined by the weighted ℓ_1-norm of the parameters

$$R(\boldsymbol{\lambda}) = \gamma||\boldsymbol{\lambda}||_1 = \gamma \sum_{i=1}^{n} |\lambda_i|$$

where γ is a regularization coefficient, and n is number of parameters. This regularizer penalizes overly large parameter values in their absolute values, and tends to force a subset of the parameters to be exactly zero at the optimum. This fact leads to a natural integration of regularization into incremental feature selection as follows: Assuming a tendency of the ℓ_1 regularizer to produce a large number of zero-valued parameters at the function's optimum, we start with all-zero weights, and incrementally add features to the model only if adjusting their parameters away from zero sufficiently decreases the optimization criterion. Since every non-zero weight added to the model incurs a regularizer penalty of $\gamma|\lambda_i|$, it only makes sense to add a feature to the model if this penalty is outweighed by the reduction in negative log-likelihood. Thus features considered for selection have to pass the following test:

$$\left|\frac{\partial L(\boldsymbol{\lambda})}{\partial \lambda_i}\right| > \gamma$$

This gradient test is applied to each feature and at each step the features that pass the test with maximum magnitude are added to the model. This provides both efficient and accurate estimation with large feature sets.

Work on discriminative reranking has been reported before by Och and Ney (2002), Och et al. (2004), and Shen et al. (2004). The main purpose of our reranking experiments is to have a system that can easily be adjusted to yield system variants that differ at controllable amounts. For quick experimental turnaround we selected the training and test data from sentences with 5 to 15 words, resulting in a training set of 160,000 sentences, and a development set of 2,000 sentences. The phrase-table employed was restricted to phrases of maximally 3 words, resulting in 200,000 phrases.

3 Detecting Small Result Differences by Intrinsic Evaluations Metrics

The intrinsic evaluation measures used in our experiments are the well-known BLEU (Papineni et al., 2001) and NIST (Doddington, 2002) metrics, and an F-score measure that adapts evaluation techniques from dependency-based parsing (Crouch et al., 2002) and sentence-condensation (Riezler et al., 2003) to machine translation. All of these measures

Set $c = 0$
Compute actual statistic of score differences $|S_{\mathrm{X}} - S_{\mathrm{Y}}|$ on test data
For random shuffles $r = 0, \ldots, R$
    For sentences in test set
        Shuffle variable tuples between system X and Y with probability 0.5
    Compute pseudo-statistic $|S_{\mathrm{X_r}} - S_{\mathrm{Y_r}}|$ on shuffled data
    If $|S_{\mathrm{X_r}} - S_{\mathrm{Y_r}}| \geq |S_{\mathrm{X}} - S_{\mathrm{Y}}|$
        $c++$
$p = (c + 1)/(R + 1)$
Reject null hypothesis if p is less than or equal to specified rejection level.

Figure 1: Approximate Randomization Test for Statistical Significance Testing

evaluate document similarity of SMT output against manually created reference translations. The measures differ in their focus on different entities in matching, corresponding to a focus on different aspects of translation quality.

BLEU and NIST both consider n-grams in source and reference strings as matching entities. BLEU weighs all n-grams equally whereas NIST puts more weight on n-grams that are more informative, i.e., occur less frequently. This results in BLEU favoring matches in larger n-grams, corresponding to giving more credit to correct word order. NIST weighs lower n-grams more highly, thus it gives more credit to correct lexical choice than to word order.

F-score is computed by parsing reference sentences and SMT outputs, and matching grammatical dependency relations. The reported value is the harmonic mean of precision and recall, which is defined as $(2 \times \text{ precision } \times \text{ recall })/(\text{ precision } + \text{ recall })$. Precision is the ratio of matching dependency relations to the total number of dependency relations in the parse for the system translation, and recall is the ratio of matches to the total number of dependency relations in the parse for the reference translation. The goal of this measure is to focus on aspects of meaning in measuring similarity of system translations to reference translations, and to allow for meaning-preserving word order variation.

Evaluation results for a comparison of reranking against a baseline model that only includes features corresponding to the 8 system components are shown in Table 1. Since the task is a comparison of system variants for development, all results are reported on the development set of 2,000 examples of length 5-15. The reranking model achieves an increase in NIST score of .15 units, whereas BLEU and F-score decrease by .3% and .2% respectively. However, as measured by the statistical significance tests described below, the differences in BLEU and F-scores are not statistically significant with p-values exceeding the standard rejection level of $.05$. In contrast, the differences in NIST score are highly significant. These findings correspond to results reported in Zhang et al. (2004) showing a higher sensitivity of NIST versus BLEU to small result differences. Taking also the results from F-score matching in account, we can conclude that similarity measures that are based on matching more complex entities (such as BLEU's higher n-grams or F's grammatical relations) are not as sensitive to small result differences as scoring techniques that are able to distinguish models by matching simpler entities (such as NIST's focus on lexical choice). Furthermore, we get an indication that differences of .3% in BLEU score or .2% in F-score might not be large enough to conclude statistical significance of result differences. This leads to questions of power and accuracy of the employed statistical significance tests which will be addressed in the next section.

4 Assessing Statistical Significance of Small Result Differences

The bootstrap method is an example for a computer-intensive statistical hypothesis test (see, e.g., Noreen (1989)). Such tests are designed to assess result differences with respect to a test statistic in cases where the sampling distribution of the test statistic

Set $c = 0$
Compute actual statistic of score differences $|S_{\mathrm{X}} - S_{\mathrm{Y}}|$ on test data
Calculate sample mean $\tau_B = \frac{1}{B}\sum_{b=0}^{B} |S_{\mathrm{X_b}} - S_{\mathrm{Y_b}}|$ over bootstrap samples $b = 0, \ldots, B$
For bootstrap samples $b = 0, \ldots, B$
    Sample with replacement from variable tuples for systems X and Y for test sentences
    Compute pseudo-statistic $|S_{\mathrm{X_b}} - S_{\mathrm{Y_b}}|$ on bootstrap data
    If $|S_{\mathrm{X_b}} - S_{\mathrm{Y_b}}| - \tau_B \ (+\tau) \geq |S_{\mathrm{X}} - S_{\mathrm{Y}}|$
      $c++$
$p = (c + 1)/(B + 1)$
Reject null hypothesis if p is less than or equal to specified rejection level.

Figure 2: Bootstrap Test for Statistical Significance Testing

is unknown. Comparative evaluations of outputs of SMT systems according to test statistics such as differences in BLEU, NIST, or F-score are examples of this situation. The attractiveness of computer-intensive significance tests such as the bootstrap or the approximate randomization method lies in their power and simplicity. As noted in standard textbooks such as Cohen (1995) or Noreen (1989) such tests are as powerful as parametric tests when parametric assumptions are met and they outperform them when parametric assumptions are violated. Because of their generality and simplicity they are also attractive alternatives to conventional non-parametric tests (see, e.g., Siegel (1988)). The power of these tests lies in the fact that they answer only a very simple question without making too many assumptions that may not be met in the experimental situation. In case of the approximate randomization test, only the question whether two samples are related to each other is answered, without assuming that the samples are representative of the populations from which they were drawn. The bootstrap method makes exactly this one assumption. This makes it formally possible to draw inferences about population parameters for the bootstrap, but not for approximate randomization. However, if the goal is to assess statistical significance of a result difference between two systems the approximate randomization test provides the desired power and accuracy whereas the bootstrap's advantage to draw inferences about population parameters comes at the price of reduced accuracy. Noreen summarizes this shortcoming of the bootstrap technique as follows: "The principal disadvantage of [the bootstrap] method is that the null hypothesis may be rejected because the shape of the sampling distribution is not well-approximated by the shape of the bootstrap sampling distribution rather than because the expected value of the test statistic differs from the value that is hypothesized."(Noreen (1989), p. 89). Below we describe these two test procedures in more detail, and compare them in our experimental setup.

4.1 Approximate Randomization

An excellent introduction to the approximate randomization test is Noreen (1989). Applications of this test to natural language processing problems can be found in Chinchor et al. (1993).

In our case of assessing statistical significance of result differences between SMT systems, the test statistic of interest is the absolute value of the difference in BLEU, NIST, or F-scores produced by two systems on the same test set. These test statistics are computed by accumulating certain count variables over the sentences in the test set. For example, in case of BLEU and NIST, variables for the length of reference translations and system translations, and for n-gram matches and n-gram counts are accumulated over the test corpus. In case of F-score, variable tuples consisting of the number of dependency-relations in the parse for the system translation, the number of dependency-relations in the parse for the reference translation, and the number of matching dependency-relations between system and reference parse, are accumulated over the test set.

Under the null hypothesis, the compared systems are not different, thus any variable tuple produced by one of the systems could have been produced just as

Table 2: NIST scores for equivalent systems under bootstrap and approximate randomization tests.

compared systems	1:2	1:3	1:4	1:5	1:6
NIST difference	.031	.032	.029	.028	.036
approxrand p-value	.03	.025	.05	.079	.028
bootstrap p-value	.014	.013	.028	.039	.013

likely by the other system. So shuffling the variable tuples between the two systems with equal probability, and recomputing the test statistic, creates an approximate distribution of the test statistic under the null hypothesis. For a test set of S sentences there are 2^S different ways to shuffle the variable tuples between the two systems. Approximate randomization produce shuffles by random assignments instead of evaluating all 2^S possible assignments. Significance levels are computed as the percentage of trials where the pseudo statistic, i.e., the test statistic computed on the shuffled data, is greater than or equal to the actual statistic, i.e., the test statistic computed on the test data. A sketch of an algorithm for approximate randomization testing is given in Fig. 1.

4.2 The Bootstrap

An excellent introduction to the technique is the textbook by Efron and Tibshirani (1993). In contrast to approximate randomization, the bootstrap method makes the assumption that the sample is a representative "proxy" for the population. The shape of the sampling distribution is estimated by repeatedly sampling (with replacement) from the sample itself.

A sketch of a procedure for bootstrap testing is given in Fig. 2. First, the test statistic is computed on the test data. Then, the sample mean of the pseudo statistic is computed on the bootstrapped data, i.e., the test statistic is computed on bootstrap samples of equal size and averaged over bootstrap samples.

In order to compute significance levels based on the bootstrap sampling distribution, we employ the "shift" method described in Noreen (1989). Here it is assumed that the sampling distribution of the null hypothesis and the bootstrap sampling distribution have the same shape but a different location. The location of the bootstrap sampling distribution is shifted so that it is centered over the location where the null hypothesis sampling distribution should be centered. This is achieved by subtracting from each value of the pseudo-statistic its expected value τ_B and then adding back the expected value τ of the test statistic under the null hypothesis. τ_B can be estimated by the sample mean of the bootstrap samples; τ is 0 under the null hypothesis. Then, similar to the approximate randomization test, significance levels are computed as the percentage of trials where the (shifted) pseudo statistic is greater than or equal to the actual statistic.

4.3 Power vs. Type I Errors

In order to evaluate accuracy of the bootstrap and the approximate randomization test, we conduct an experimental evaluation of type-I errors of both bootstrap and approximate randomization on real data. Type-I errors indicate failures to reject the null hypothesis when it is true. We construct SMT system variants that are essentially equal but produce superficially different results. This can be achieved by constructing reranking variants that differ in the redundant features that are included in the models, but are similar in the number and kind of selected features. The results of this experiment are shown in Table 2. System 1 does not include irrelevant features, whereas systems 2-6 were constructed by adding a slightly different number of features in each step, but resulted in the same number of selected features. Thus competing features bearing the same information are exchanged in different models, yet overall the same information is conveyed by slightly different feature sets. The results of Table 2 show that the bootstrap method yields p-values $< .015$ in 3 out of 5 pairwise comparisons whereas the approximate randomization test yields p-values $\geq .025$ in all cases. Even if the true p-value is unknown, we can say that the approximate randomization test estimates p-values more conservatively than the bootstrap, thus increasing the likelihood of type-I error for the bootstrap test. For a restrictive significance level of 0.15, which is motivated below for multiple

comparisons, the bootstrap would assess statistical significance in 3 out of 5 cases whereas statistical significance would not be assessed under approximate randomization. Assuming equivalence of the compared system variants, these assessments would count as type-I errors.

4.4 The Multiplicity Problem

In the experiment on type-I error described above, a more stringent rejection level than the usual $.05$ was assumed. This was necessary to circumvent a common pitfall in significance testing for k-fold pairwise comparisons. Following the argumentation given in Cohen (1995), the probability of randomly assessing statistical significance for result differences in k-fold pairwise comparisons grows exponentially in k. Recall that for a pairwise comparison of systems, specifying that $p < .05$ means that the probability of incorrectly rejecting the null hypothesis that the systems are not different be less than $.05$. Caution has to be exercised in k-fold pairwise comparisons: For a probability p_c of incorrectly rejecting the null hypothesis in a specific pairwise comparison, the probability p_e of at least once incorrectly rejecting this null hypothesis in an experiment involving k pairwise comparisons is

$$p_e \approx 1 - (1 - p_c)^k$$

For large values of k, the probability of concluding result differences incorrectly at least once is undesirably high. For example, in benchmark testing of 15 systems, $15(15-1)/2 = 105$ pairwise comparisons will have to be conducted. At a per-comparison rejection level $p_c = .05$ this results in an experimentwise error $p_e = .9954$, i.e., the probability of at least one spurious assessment of significance is $1 - (1 - .05)^{105} = .9954$. One possibility to reduce the likelihood that one ore more of differences assessed in pairwise comparisons is spurious is to run the comparisons at a more stringent per-comparison rejection level. Reducing the per-comparison rejection level p_c until an experimentwise error rate p_e of a standard value, e.g., $.05$, is achieved, will favor p_e over p_c. In the example of 5 pairwise comparisons described above, a per-comparison error rate $p_c = .015$ was sufficient to achieve an experimentwise error rate $p_e \approx .07$. In many cases this technique would require to reduce p_c to the point where a result difference has to be unrealistically large to be significant. Here conventional tests for post-hoc comparisons such as the Scheffé or Tukey test have to be employed (see Cohen (1995), p. 185ff.).

5 Conclusion

Situations where a researcher has to deal with subtle differences between systems are common in system development and large benchmark tests. We have shown that it is useful in such situations to trade in expressivity of evaluation measures for sensitivity. For MT evaluation this means that recording differences in lexical choice by the NIST measure is more useful than failing to record differences by employing measures such as BLEU or F-score that incorporate aspects of fluency and meaning adequacy into MT evaluation. Similarly, in significance testing, it is useful to trade in the possibility to draw inferences about the sampling distribution for accuracy and power of the test method. We found experimental evidence confirming textbook knowledge about reduced accuracy of the bootstrap test compared to the approximate randomization test. Lastly, we pointed out a well-known problem of randomly assessing significance in multiple pairwise comparisons. Taking these findings together, we recommend for multiple comparisons of subtle differences to combine the NIST score for evaluation with the approximate randomization test for significance testing, at more stringent rejection levels than is currently standard in the MT literature.

References

Nancy Chinchor, Lynette Hirschman, and David D. Lewis. 1993. Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3). *Computational Linguistics*, 19(3):409–449.

Paul R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. The MIT Press, Cambridge, MA.

Richard Crouch, Ronald M. Kaplan, Tracy H. King, and Stefan Riezler. 2002. A comparison of evaluation metrics for a broad-coverage stochastic parser. In *Proceedings of the "Beyond PARSEVAL" Workshop at the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Spain.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence

statistics. In *Proceedings of the ARPA Workshop on Human Language Technology*.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Kevin Knight and Yaser Al-Onaizan. 1999. A primer on finite-state software for natural language processing. Technical report, USC Information Sciences Institute, Marina del Rey, CA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, Edmonton, Cananda.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Technical report, USC Information Sciences Institute, Marina del Rey, CA.

Philipp Koehn. 2004a. PHARAOH. a beam search decoder for phrase-based statistical machine translation models. user manual. Technical report, USC Information Sciences Institute, Marina del Rey, CA.

Philipp Koehn. 2004b. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, Barcelona, Spain.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL'04)*, Boston, MA.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley, New York.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the 1999 Conference on Empirical Methods in Natural Language Processing (EMNLP'99)*.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Ketherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL'04)*, Boston, MA.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, Edmonton, Cananda.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report IBM Research Division Technical Report, RC22176 (W0190-022), Yorktown Heights, N.Y.

Stefan Riezler and Alexander Vasserman. 2004. Incremental feature selection and ℓ_1 regularization for relaxed maximum-entropy modeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, Barcelona, Spain.

Stefan Riezler, Tracy H. King, Richard Crouch, and Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, Edmonton, Cananda.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL'04)*, Boston, MA.

Sidney Siegel. 1988. *Nonparametric Statistics for the Behavioral Sciences. Second Edition*. MacGraw-Hill, Boston, MA.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments

Satanjeev Banerjee
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
banerjee+@cs.cmu.edu

Alon Lavie
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
alavie@cs.cmu.edu

Abstract

We describe METEOR, an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations. Unigrams can be matched based on their surface forms, stemmed forms, and meanings; furthermore, METEOR can be easily extended to include more advanced matching strategies. Once all generalized unigram matches between the two strings have been found, METEOR computes a score for this matching using a combination of unigram-precision, unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference. We evaluate METEOR by measuring the correlation between the metric scores and human judgments of translation quality. We compute the Pearson R correlation value between its scores and human quality assessments of the LDC TIDES 2003 Arabic-to-English and Chinese-to-English datasets. We perform segment-by-segment correlation, and show that METEOR gets an R correlation value of 0.347 on the Arabic data and 0.331 on the Chinese data. This is shown to be an improvement on using simply unigram-precision, unigram-recall and their harmonic F1 combination. We also perform experiments to show the relative contributions of the various mapping modules.

1 Introduction

Automatic Metrics for machine translation (MT) evaluation have been receiving significant attention in the past two years, since IBM's BLEU metric was proposed and made available (Papineni et al 2002). BLEU and the closely related NIST metric (Doddington, 2002) have been extensively used for comparative evaluation of the various MT systems developed under the DARPA TIDES research program, as well as by other MT researchers. The utility and attractiveness of automatic metrics for MT evaluation has consequently been widely recognized by the MT community. Evaluating an MT system using such automatic metrics is much faster, easier and cheaper compared to human evaluations, which require trained bilingual evaluators. In addition to their utility for comparing the performance of different systems on a common translation task, automatic metrics can be applied on a frequent and ongoing basis during system development, in order to guide the development of the system based on concrete performance improvements.

Evaluation of Machine Translation has traditionally been performed by humans. While the main criteria that should be taken into account in assessing the quality of MT output are fairly intuitive and well established, the overall task of MT evaluation is both complex and task dependent. MT evaluation has consequently been an area of significant research in itself over the years. A wide range of assessment measures have been proposed, not all of which are easily quantifiable. Recently developed frameworks, such as FEMTI (King et al, 2003), are attempting to devise effective platforms for combining multi-faceted measures for MT evaluation in effective and user-adjustable ways. While a single one-dimensional numeric metric cannot hope to fully capture all aspects of MT

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, June 2005.

evaluation, such metrics are still of great value and utility.

In order to be both effective and useful, an automatic metric for MT evaluation has to satisfy several basic criteria. The primary and most intuitive requirement is that the metric have very high correlation with quantified human notions of MT quality. Furthermore, a good metric should be as sensitive as possible to differences in MT quality between different systems, and between different versions of the same system. The metric should be consistent (same MT system on similar texts should produce similar scores), reliable (MT systems that score similarly can be trusted to perform similarly) and general (applicable to different MT tasks in a wide range of domains and scenarios). Needless to say, satisfying all of the above criteria is extremely difficult, and all of the metrics that have been proposed so far fall short of adequately addressing most if not all of these requirements. Nevertheless, when appropriately quantified and converted into concrete test measures, such requirements can set an overall standard by which different MT evaluation metrics can be compared and evaluated.

In this paper, we describe METEOR[1], an automatic metric for MT evaluation which we have been developing. METEOR was designed to explicitly address several observed weaknesses in IBM's BLEU metric. It is based on an explicit word-to-word matching between the MT output being evaluated and one or more reference translations. Our current matching supports not only matching between words that are identical in the two strings being compared, but can also match words that are simple morphological variants of each other (i.e. they have an identical stem), and words that are synonyms of each other. We envision ways in which this strict matching can be further expanded in the future, and describe these at the end of the paper. Each possible matching is scored based on a combination of several features. These currently include unigram-precision, unigram-recall, and a direct measure of how out-of-order the words of the MT output are with respect to the reference. The score assigned to each individual sentence of MT output is derived from the best scoring match among all matches over all reference translations. The maximal-scoring matching is then also used in order to calculate an aggregate score for the MT system over the entire test set. Section 2 describes the metric in detail, and provides a full example of the matching and scoring.

In previous work (Lavie et al., 2004), we compared METEOR with IBM's BLEU metric and it's derived NIST metric, using several empirical evaluation methods that have been proposed in the recent literature as concrete means to assess the level of correlation of automatic metrics and human judgments. We demonstrated that METEOR has significantly improved correlation with human judgments. Furthermore, our results demonstrated that recall plays a more important role than precision in obtaining high-levels of correlation with human judgments. The previous analysis focused on correlation with human judgments at the *system level*. In this paper, we focus our attention on improving correlation between METEOR score and human judgments at the *segment level*. High-levels of correlation at the segment level are important because they are likely to yield a metric that is sensitive to minor differences between systems and to minor differences between different versions of the same system. Furthermore, current levels of correlation at the sentence level are still rather low, offering a very significant space for improvement. The results reported in this paper demonstrate that all of the individual components included within METEOR contribute to improved correlation with human judgments. In particular, METEOR is shown to have statistically significant better correlation compared to unigram-precision, unigram-recall and the harmonic F1 combination of the two.

We are currently in the process of exploring several further enhancements to the current METEOR metric, which we believe have the potential to significantly further improve the sensitivity of the metric and its level of correlation with human judgments. Our work on these directions is described in further detail in Section 4.

2 The METEOR Metric

2.1 Weaknesses in BLEU Addressed in METEOR

The main principle behind IBM's BLEU metric (Papineni et al, 2002) is the measurement of the

[1] METEOR: **M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering

overlap in unigrams (single words) and higher order n-grams of words, between a translation being evaluated and a set of one or more reference translations. The main component of BLEU is *n-gram precision*: the proportion of the matched n-grams out of the total number of n-grams in *the evaluated translation.* Precision is calculated separately for each n-gram order, and the precisions are combined via a geometric averaging. BLEU does not take *recall* into account directly. Recall – the proportion of the matched n-grams out of the total number of n-grams in *the reference translation*, is extremely important for assessing the quality of MT output, as it reflects to what degree the translation covers the entire content of the translated sentence. BLEU does not use recall because the notion of recall is unclear when matching simultaneously against a set of reference translations (rather than a single reference). To compensate for recall, BLEU uses a Brevity Penalty, which penalizes translations for being "too short". The NIST metric is conceptually similar to BLEU in most aspects, including the weaknesses discussed below.

BLEU and NIST suffer from several weaknesses, which we attempt to address explicitly in our proposed METEOR metric:

The Lack of Recall: We believe that the fixed brevity penalty in BLEU does not adequately compensate for the lack of recall. Our experimental results strongly support this claim.

Use of Higher Order N-grams: Higher order N-grams are used in BLEU as an indirect measure of a translation's level of grammatical well-formedness. We believe an explicit measure for the level of grammaticality (or word order) can better account for the importance of grammaticality as a factor in the MT metric, and result in better correlation with human judgments of translation quality.

Lack of Explicit Word-matching Between Translation and Reference: N-gram counts don't require an explicit word-to-word matching, but this can result in counting incorrect "matches", particularly for common function words.

Use of Geometric Averaging of N-grams: Geometric averaging results in a score of "zero" whenever one of the component n-gram scores is zero. Consequently, BLEU scores at the sentence (or segment) level can be meaningless. Although BLEU was intended to be used only for aggregate counts over an entire test-set (and not at the sentence level), scores at the sentence level can be useful indicators of the quality of the metric. In experiments we conducted, a modified version of BLEU that uses equal-weight arithmetic averaging of n-gram scores was found to have better correlation with human judgments.

2.2 The METEOR Metric

METEOR was designed to explicitly address the weaknesses in BLEU identified above. It evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation. If more than one reference translation is available, the given translation is scored against each reference independently, and the best score is reported. This is discussed in more detail later in this section.

Given a pair of translations to be compared (a system translation and a reference translation), METEOR creates an *alignment* between the two strings. We define an alignment as a mapping between unigrams, such that every unigram in each string maps to zero or one unigram in the other string, and to no unigrams in the same string. Thus in a given alignment, a single unigram in one string cannot map to more than one unigram in the other string. This alignment is incrementally produced through a series of *stages,* each stage consisting of two distinct phases.

In the first phase an external module lists all the possible unigram mappings between the two strings. Thus, for example, if the word "computer" occurs once in the system translation and twice in the reference translation, the external module lists two possible unigram mappings, one mapping the occurrence of "computer" in the system translation to the first occurrence of "computer" in the reference translation, and another mapping it to the second occurrence. Different modules map unigrams based on different criteria. The "exact" module maps two unigrams if they are exactly the same (e.g. "computers" maps to "computers" but not "computer"). The "porter stem" module maps two unigrams if they are the same *after* they are stemmed using the Porter stemmer (e.g.: "computers" maps to both "computers" and to "computer"). The "WN synonymy" module maps two unigrams if they are synonyms of each other.

In the second phase of each stage, the largest subset of these unigram mappings is selected such

that the resulting set constitutes an *alignment* as defined above (that is, each unigram must map to at most one unigram in the other string). If more than one subset constitutes an alignment, and also has the same cardinality as the largest set, METEOR selects that set that has the least number of unigram mapping *crosses*. Intuitively, if the two strings are typed out on two rows one above the other, and lines are drawn connecting unigrams that are mapped to each other, each line crossing is counted as a "unigram mapping cross". Formally, two unigram mappings (t_i, r_j) and (t_k, r_l) (where t_i and t_k are unigrams in the system translation mapped to unigrams r_j and r_l in the reference translation respectively) are said to *cross* if and only if the following formula evaluates to a negative number:

$(pos(t_i) - pos(t_k)) * (pos(r_j) - pos(r_l))$

where $pos(t_x)$ is the numeric position of the unigram t_x in the system translation string, and $pos(r_y)$ is the numeric position of the unigram r_y in the reference string. For a given alignment, every pair of unigram mappings is evaluated as a cross or not, and the alignment with the least total crosses is selected in this second phase. Note that these two phases together constitute a variation of the algorithm presented in (Turian et al, 2003).

Each stage only maps unigrams that have not been mapped to any unigram in any of the preceding stages. Thus the order in which the stages are run imposes different priorities on the mapping modules employed by the different stages. That is, if the first stage employs the "exact" mapping module and the second stage employs the "porter stem" module, METEOR is effectively preferring to first map two unigrams based on their surface forms, and performing the stemming only if the surface forms do not match (or if the mapping based on surface forms was too "costly" in terms of the total number of crosses). Note that METEOR is flexible in terms of the number of stages, the actual external mapping module used for each stage, and the order in which the stages are run. By default the first stage uses the "exact" mapping module, the second the "porter stem" module and the third the "WN synonymy" module. In section 4 we evaluate each of these configurations of METEOR.

Once all the stages have been run and a final alignment has been produced between the system translation and the reference translation, the METEOR score for this pair of translations is computed as follows. First unigram precision (P) is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the system translation. Similarly, unigram recall (R) is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the reference translation. Next we compute *Fmean* by combining the precision and recall via a harmonic-mean (van Rijsbergen, 1979) that places most of the weight on recall. We use a harmonic mean of P and 9R. The resulting formula used is:

$$Fmean = \frac{10PR}{R + 9P}$$

Precision, recall and Fmean are based on unigram matches. To take into account longer matches, METEOR computes a *penalty* for a given alignment as follows. First, all the unigrams in the system translation that are mapped to unigrams in the reference translation are grouped into the fewest possible number of *chunks* such that the unigrams in each chunk are in adjacent positions in the system translation, and are also mapped to unigrams that are in adjacent positions in the reference translation. Thus, the longer the n-grams, the fewer the chunks, and in the extreme case where the entire system translation string matches the reference translation there is only one chunk. In the other extreme, if there are no bigram or longer matches, there are as many chunks as there are unigram matches. The penalty is then computed through the following formula:

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

For example, if the system translation was "the president spoke to the audience" and the reference translation was "the president then spoke to the audience", there are two chunks: "the president" and "spoke to the audience". Observe that the penalty increases as the number of chunks increases to a maximum of *0.5*. As the number of chunks goes to 1, penalty decreases, and its lower bound is decided by the number of unigrams matched. The parameters if this penalty function were determined based on some experimentation with de-

veopment data, but have not yet been trained to be optimal.

Finally, the METEOR *Score* for the given alignment is computed as follows:

$$Score = Fmean * (1 - Penalty)$$

This has the effect of *reducing* the Fmean by the maximum of 50% if there are *no* bigram or longer matches.

For a single system translation, METEOR computes the above score for each reference translation, and then reports the best score as the score for the translation. The overall METEOR score for a system is calculated based on aggregate statistics accumulated over the entire test set, similarly to the way this is done in BLEU. We calculate aggregate precision, aggregate recall, an aggregate penalty, and then combine them using the same formula used for scoring individual segments.

3 Evaluation of the METEOR Metric

3.1. Data

We evaluated the METEOR metric and compared its performance with BLEU and NIST on the DARPA/TIDES 2003 Arabic-to-English and Chinese-to-English MT evaluation data released through the LDC as a part of the workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, at the Annual Meeting of the Association of Computational Linguistics (2005). The Chinese data set consists of 920 sentences, while the Arabic data set consists of 664 sentences. Each sentence has four reference translations. Furthermore, for 7 systems on the Chinese data and 6 on the Arabic data, every sentence translation has been assessed by two separate human judges and assigned an *Adequacy* and a *Fluency Score*. Each such score ranges from one to five (with one being the poorest grade and five the highest). For this paper, we computed a *Combined Score* for each translation by averaging the adequacy and fluency scores of the two judges for that translation. We also computed an average *System Score* for each translation system by averaging the Combined Score for all the translations produced by that system. (Note that although we refer to these data sets as the "Chinese" and the "Arabic" data sets, the MT evaluation systems analyzed in this paper only evaluate English sentences produced by translation systems by comparing them to English reference sentences).

3.2 Comparison with BLEU and NIST MT Evaluation Algorithms

In this paper, we are interested in evaluating METEOR as a metric that can evaluate translations on a sentence-by-sentence basis, rather than on a coarse grained system-by-system basis. The standard metrics – BLEU and NIST – were however designed for system level scoring, hence computing sentence level scores using BLEU or the NIST evaluation mechanism is unfair to those algorithms. To provide a point of comparison however, table 1 shows the *system level* correlation between human judgments and various MT evaluation algorithms and sub components of METEOR over the Chinese portion of the Tides 2003 dataset. Specifically, these correlation figures were obtained as follows: Using each algorithm we computed one score per Chinese system by calculating the aggregate scores produced by that algorithm for that system. We also obtained the overall human judgment for each system by averaging all the human scores for that system's translations. We then computed the Pearson correlation between these system level human judgments and the system level scores for each algorithm; these numbers are presented in table 1.

System ID	Correlation
BLEU	0.817
NIST	0.892
Precision	0.752
Recall	0.941
F1	0.948
Fmean	0.952
METEOR	0.964

Table 1: Comparison of human/METEOR correlation with BLEU and NIST/human correlations

Observe that simply using Recall as the MT evaluation metric results in a significant improvement in correlation with human judgment over both the BLEU and the NIST algorithms. These correlations further improve slightly when precision is taken into account (in the F1 measure),

when the recall is weighed more heavily than precision (in the Fmean measure) and when a penalty is levied for fragmented matches (in the main METEOR measure).

3.3 Evaluation Methodology

As mentioned in the previous section, our main goal in this paper is to evaluate METEOR and its components on their translation-by-translation level correlation with human judgment. Towards this end, in the rest of this paper, our evaluation methodology is as follows: For each system, we compute the METEOR Score for every translation produced by the system, and then compute the correlation between these *individual* scores and the human assessments (average of the adequacy and fluency scores) for the same translations. Thus we get a single *Pearson R value* for each system for which we have human assessments. Finally we average the *R* values of all the systems for each of the two language data sets to arrive at the overall average correlation for the Chinese dataset and the Arabic dataset. This number ranges between -1.0 (completely negatively correlated) to +1.0 (completely positively correlated).

We compare the correlation between human assessments and METEOR Scores produced above with that between human assessments and precision, recall and Fmean scores to show the advantage of the various components in the METEOR scoring function. Finally we run METEOR using different mapping modules, and compute the correlation as described above for each configuration to show the effect of each unigram mapping mechanism.

3.4 Correlation between METEOR Scores and Human Assessments

System ID	Correlation
ame	0.331
ara	0.278
arb	0.399
ari	0.363
arm	0.341
arp	0.371
Average	**0.347**

Table 2: Correlation between METEOR Scores and Human Assessments for the Arabic Dataset

We computed sentence by sentence correlation between METEOR Scores and human assessments (average of adequacy and fluency scores) for each translation for every system. Tables 2 and 3 show the Pearson R correlation values for each system, as well as the average correlation value per language dataset.

System ID	Correlation
E09	0.385
E11	0.299
E12	0.278
E14	0.307
E15	0.306
E17	0.385
E22	0.355
Average	**0.331**

Table 3: Correlation between METEOR Scores and Human Assessments for the Chinese Dataset

3.5 Comparison with Other Metrics

We computed translation by translation correlations between human assessments and other metrics besides the METEOR score, namely precision, recall and Fmean. Tables 4 and 5 show the correlations for the various scores.

Metric	Correlation
Precision	0.287
Recall	0.334
Fmean	0.340
METEOR	**0.347**

Table 4: Correlations between human assessments and precision, recall, Fmean and METEOR Scores, averaged over systems in the Arabic dataset

Metric	Correlation
Precision	0.286
Recall	0.320
Fmean	0.327
METEOR	**0.331**

Table 5: Correlations between human assessments and precision, recall, Fmean and METEOR Scores, averaged over systems in the Chinese dataset

We observe that recall by itself correlates with human assessment much better than precision, and that combining the two using the Fmean formula

described above results in further improvement. By penalizing the Fmean score using the chunk count we get some further marginal improvement in correlation.

3.6 Comparison between Different Mapping Modules

To observe the effect of various unigram mapping modules on the correlation between the METEOR score and human assessments, we ran METEOR with different sequences of stages with different mapping modules in them. In the first experiment we ran METEOR with only one stage that used the "exact" mapping module. This module matches unigrams only if their surface forms match. (This module does not match unigrams that belong to a list of "stop words" that consist mainly of function words). In the second experiment we ran METEOR with two stages, the first using the "exact" mapping module, and the second the "Porter" mapping module. The Porter mapping module matches two unigrams to each other if they are identical after being passed through the Porter stemmer. In the third experiment we replaced the Porter mapping module with the WN-Stem mapping module. This module maps two unigrams to each other if they share the same *base form* in WordNet. This can be thought of as a different kind of stemmer – the difference from the Porter stemmer is that the word stems are actual words when stemmed through WordNet in this manner. In the last experiment we ran METEOR with three stages, the first two using the exact and the Porter modules, and the third the WN-Synonymy mapping module. This module maps two unigrams together if at least one sense of each word belongs to the same synset in WordNet. Intuitively, this implies that at least one sense of each of the two words represent the same concept. This can be thought of as a poor-man's synonymy detection algorithm that does not disambiguate the words being tested for synonymy. Note that the METEOR scores used to compute correlations in the other tables (1 through 4) used exactly this sequence of stages.

Tables 6 and 7 show the correlations between METEOR scores produced in each of these experiments and human assessments for both the Arabic and the Chinese datasets. On both data sets, adding either stemming modules to simply using the exact matching improves correlations. Some further improvement in correlation is produced by adding the synonymy module.

Mapping module sequence used (Arabic)	Correlation
Exact	0.312
Exact, Porter	0.329
Exact, WN-Stem	0.330
Exact, Porter, WN-Synonym	0.347

Table 6: Comparing correlations produced by different module stages on the Arabic dataset.

Mapping module sequence used (Chinese)	Correlation
Exact	0.293
Exact, Porter	0.318
Exact, WN-Stem	0.312
Exact, Porter, WN-Synonym	0.331

Table 7: Comparing correlations produced by different module stages, on the Chinese dataset

3.7 Correlation using Normalized Human Assessment Scores

One problem with conducting correlation experiments with human assessment scores at the sentence level is that the human scores are noisy – that is, the levels of agreement between human judges on the actual sentence level assessment scores is not extremely high. To partially address this issue, the human assessment scores were normalized by a group at the MITRE Corporation. To see the effect of this noise on the correlation, we computed the correlation between the METEOR Score (computed using the stages used in the 4th experiment in section 7 above) and both the raw human assessments as well as the normalized human assessments.

	Arabic Dataset	Chinese Dataset
Raw human assessments	0.347	0.331
Normalized human assessments	0.403	0.365

Table 8: Comparing correlations between METEOR Scores and both raw and normalized human assessments

Table 8 shows that indeed METEOR Scores correlate better with normalized human assessments. In other words, the noise in the human assessments hurts the correlations between automatic scores and human assessments.

4 Future Work

The METEOR metric we described and evaluated in this paper, while already demonstrating great promise, is still relatively simple and naïve. We are in the process of enhancing the metric and our experimentation in several directions:

Train the Penalty and Score Formulas on Data: The formulas for Penalty and METEOR score were manually crafted based on empirical tests on a separate set of development data. However, we plan to optimize the formulas by *training* them on a separate data set, and choosing that formula that best correlates with human assessments on the training data.

Use Semantic Relatedness to Map Unigrams: So far we have experimented with exact mapping, stemmed mapping and synonymy mapping between unigrams. Our next step is to experiment with different measures of semantic relatedness to match unigrams that have a related meaning, but are not quite synonyms of each other.

More Effective Use of Multiple Reference Translations: Our current metric uses multiple reference translations in a weak way: we compare the translation with each reference separately and select the reference with the best match. This was necessary in order to incorporate recall in our metric, which we have shown to be highly advantageous. As our matching approach improves, the need for multiple references for the metric may in fact diminish. Nevertheless, we are exploring ways in which to improve our matching against multiple references. Recent work by (Pang et al, 2003) provides the mechanism for producing semantically meaningful additional "synthetic" references from a small set of real references. We plan to explore whether using such synthetic references can improve the performance of our metric.

Weigh Matches Produced by Different Modules Differently: Our current multi-stage approach prefers metric imposes a priority on the different matching modules. However, once all the stages have been run, unigrams mapped through different mapping modules are treated the same. Another approach to treating different mappings differently is to apply different *weights* to the mappings produced by different mapping modules. Thus "computer" may match "computer" with a score of 1, "computers" with a score of 0.8 and "workstation" with a score of 0.3. As future work we plan to develop a version of METEOR that uses such weighting schemes.

Acknowledgements

We acknowledge Kenji Sagae and Shyamsundar Jayaraman for their work on the METEOR system. We also wish to thank John Henderson and William Morgan from MITRE for providing us with the normalized human judgment scores used for this work.

References

George Doddington. 2002. *Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics.* In Proceedings of 2nd Human Language Technologies Conference (HLT-02). San Diego, CA. pp. 128-132.

Margaret King, Andrei Popescu-Belis and Eduard Hovy. 2003. *FEMTI: Creating and Using a Framework for MT Evaluation.* In Proceedings of MT Summit IX, New Orleans, LA. Sept. 2003. pp. 224-231.

Alon Lavie, Kenji Sagae and Shyamsundar Jayaraman, 2004. *The Significance of Recall in Automatic Metrics for MT Evaluation*. In Proceedings of AMTA-2004, Washington DC. September 2004.

Bo Pang, Kevin Knight and Daniel Marcu. 2003. *Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences.* In Proceedings of HLT-NAACL 2003. Edmonton, Canada. May 2003.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA. July 2002. pp. 311-318.

Joseph P. Turian, Luke Shen and I. Dan Melamed. 2003. *Evaluation of Machine Translation and its Evaluation.* In Proceedings of MT Summit IX, New Orleans, LA. Sept. 2003. pp. 386-393.

C. van Rijsbergen. 1979. *Information Retrieval.* Butterworths. London, England. 2nd Edition.

Author Index