Interpretation and Transformation for Abstracting Conversations

Gabriel Murray gabrielm@cs.ubc.ca
Giuseppe Carenini carenini@cs.ubc.ca
Raymond Ng rng@cs.ubc.ca
Department of Computer Science, University of British Columbia
Vancouver, Canada

Abstract

We address the challenge of automatically abstracting conversations such as face-to-face meetings and emails. We focus here on the stages of *interpretation*, where sentences are mapped to a conversation ontology, and *transformation*, where the summary content is selected. Our approach is fully developed and tested on meeting speech, and we subsequently explore its application to email conversations.

1 Introduction

The dominant approach to the challenge of automatic summarization has been *extraction*, where informative sentences in a document are identified and concatenated to form a condensed version of the original document. Extractive summarization has been popular at least in part because it is a binary classification task that lends itself well to machine learning techniques, and does not require a natural language generation (NLG) component. There is evidence that human abstractors at times use sentences from the source documents nearly verbatim in their own summaries, justifying this approach to some extent (Kupiec et al., 1995). Extrinsic evaluations have also shown that, while extractive summaries may be less coherent than human abstracts, users still find them to be valuable tools for browsing documents (He et al., 1999; McKeown et al., 2005; Murray et al., 2008).

However, these same evaluations also indicate that concise abstracts are generally preferred by users and lead to higher objective task scores. The limitation of a cut-and-paste summary is that the end-user does not know *why* the selected sentences are important; this can often only be discerned by exploring the context in which each sentence originally appeared. One possible improvement is to create *structured extracts* that represent an increased level of abstraction, where selected sentences are grouped according to phenomena such as *decisions*, *action items* and *problems*, thereby giving the user more information on why the sentences are being highlighted. For example, the sentence *Let's go with a simple chip* represents a decision. An even higher level of abstraction can be provided by generating new text that synthesizes or extrapolates on the information contained in the structured summary. For example, the sentence *Sandra and Sue expressed negative opinions about the remote control design* can be coupled with extracted sentences containing these negative opinions, forming a *hybrid summary*. Our summarization system ultimately performs both types of abstraction, grouping sentences according to various sentence-level phenomena, and generating novel text that describes this content at a higher level.

In this work we describe the first two components of our abstractive summarization system. In the *interpretation* stage, sentences are mapped to nodes in a conversation ontology by utilizing classifiers relating to a variety of sentence-level phenomena such as *decisions*, *action items* and *subjective sentences*. These classifiers achieve high accuracy by using a very large feature set integrating conversation structure, lexical patterns, part-of-speech (POS) tags and character n-grams. In the *transformation* stage, we select the most informative sentences by maximizing a function based on the derived ontology mapping and the coverage of weighted entities mentioned in the conversation. This transformation component utilizes integer linear programming (ILP) and we compare its performance with several greedy selection algorithms.

We do not discuss the generation component of our summarization system in this paper. The transformation component is still ex-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 894–902,
Los Angeles, California, June 2010.

tractive in nature, but the sentences that are selected in the transformation stage correspond to objects in the ontology and the properties linking them. Specifically, these are triples of the form $< participant, relation, entity >$ where a $participant$ is a person in the conversation, an $entity$ is an item under discussion, and a $relation$ such as *positive opinion* or *action item* links the two. This intermediate output enables us to create structured extracts as described above, with the triples also acting as input to the downstream NLG component.

We have tested our approach in summarization experiments on both meeting and email conversations, where the quality of a sentence is measured by how effectively it conveys information in a model abstract summary according to human annotators. On meetings the ILP approach consistently outperforms several greedy summarization methods. A key finding is that emails exhibit markedly varying conversation structures, and the email threads yielding the best summarization results are those that are structured similarly to meetings. Other email conversation structures are less amenable to the current treatment and require further investigation and possibly domain adaptation.

2 Related Research

The view that summarization consists of stages of *interpretation*, *transformation* and *generation* was laid out by Sparck-Jones (1999). Popular approaches to text extraction essentially collapse interpretation and transformation into one step, with generation either being ignored or consisting of post-processing techniques such as sentence compression (Knight and Marcu, 2000; Clarke and Lapata, 2006) or sentence merging (Barzilay and McKeown, 2005). In contrast, in this work we clearly separate interpretation from transformation.

The most relevant research to ours is by Kleinbauer et al. (2007), similarly focused on meeting abstraction. They create an ontology for the AMI scenario meeting corpus (Carletta et al., 2005), described in Section 5.1. The system uses topic segments and topic labels, and for each topic segment in the meeting a sentence is generated that describes the most frequently mentioned content items in that topic. Our systems differ in two major respects: their summarization process uses human gold-standard annotations of topic segments, topic labels and content items from the ontology, while our summarizer is fully automatic; and the ontology used by Kleinbauer et al. is specific not just to meetings but to the AMI scenario meetings, while our ontology applies to conversations in general.

While the work by Kleinbauer et al. is among the earliest research on abstracting multi-party dialogues, much attention in recent years has been paid to extractive summarization of such conversations, including meetings (Galley, 2006), emails (Rambow et al., 2004; Carenini et al., 2007), telephone conversations (Zhu and Penn, 2006) and internet relay chats (Zhou and Hovy, 2005).

Recent research has addressed the challenges of detecting decisions (Hsueh et al., 2007), action items (Purver et al., 2007; Murray and Renals, 2008) and subjective sentences (Raaijmakers et al., 2008). In our work we perform all of these tasks but rely on general conversational features without recourse to meeting-specific or email-specific features.

Our approach to transformation is an adaptation of an ILP sentence selection algorithm described by Xie et al. (2009). We describe both ILP approaches in Section 4.

3 Interpretation - Ontology Mapping

Source document interpretation in our system relies on a simple conversation ontology. The ontology is written in OWL/RDF and contains two core upper-level classes: Participant and Entity. When additional information is available about participant roles in a given domain, Participant subclasses such as ProjectManager can be utilized. The ontology also contains six properties that express relations between the participants and the entities. For example, the following snippet of the ontology indicates that *hasActionItem* is a relationship between a meeting participant (the property domain) and a discussed entity (the property range).

```
<owl:ObjectProperty rdf:ID="hasActionItem">
  <rdfs:domain rdf:resource="#Participant"/>
  <rdfs:range rdf:resource="#Entity"/>
</owl:ObjectProperty>
```

Similar properties exist for decisions, actions, problems, positive-subjective sentences, negative-

subjective sentences and general extractive sentences (important sentences that may not match the other categories), all connecting conversation participants and entities. The goal is to populate the ontology with participant and entity instances from a given conversation and determine their relationships. This involves identifying the important entities and classifying the sentences in which they occur as being decision sentences, action item sentences, etc.

Our current definition of entity is simple. The entities in a conversation are noun phrases with mid-range document frequency. This is similar to the definition of concept as defined by Xie et al. (Xie et al., 2009), where n-grams are weighted by *tf.idf* scores, except that we use noun phrases rather than any n-grams because we want to refer to the entities in the generated text. We use mid-range document frequency instead of *idf* (Church and Gale, 1995), where the entities occur in between 10% and 90% of the documents in the collection. In Section 4 we describe how we use the entity's term frequency to detect the most informative entities. We do not currently attempt coreference resolution for entities; recent work has investigated coreference resolution for multi-party dialogues (Muller, 2007; Gupta et al., 2007), but the challenge of resolution on such noisy data is highlighted by low accuracy (e.g. F-measure of 21.21) compared with using well-formed text (e.g. monologues).

We map sentences to our ontology's object properties by building numerous supervised classifiers trained on labeled decision sentences, action sentences, etc. A general extractive classifier is also trained on sentences simply labeled as important. After predicting these sentence-level properties, we consider a participant to be linked to an entity if the participant mentioned the entity in a sentence in which one of these properties is predicted. We give a specific example of the ontology mapping using this excerpt from the AMI corpus:

1. A: And you two are going to work together on a *prototype* using *modelling clay*.
2. A: You'll get *specific instructions* from your *personal coach*.
3. C: Cool.
4. A: Um did we decide on a *chip*?
5. A: Let's go with a *simple chip*.

Example entities are italicized. Sentences 1 and 2 are classified as action items. Sentence 3 is classified as positive-subjective, but because it contains no entities, no $< participant, relation, entity >$ triple can be added to the ontology. Sentence 4 is classified as a decision sentence, and Sentence 5 is both a decision sentence and a positive-subjective sentence (because the participant is advocating a particular position). The ontology is populated by adding all of the sentence entities as instances of the Entity class, all of the participants as instances of the Participant class, and adding $< participant, relation, entity >$ triples for Sentences 1, 2, 4 and 5. For example, Sentence 5 results in the following two triples being added to the ontology:

```
<ProjectManager rdf:ID="participant-A">
<hasDecision rdf:resource="#simple-chip"/>
</ProjectManager>
```

```
<ProjectManager rdf:ID="participant-A">
<hasPos rdf:resource="#simple-chip"/>
</ProjectManager>
```

Elements in the ontology are associated with linguistic annotations used by the generation component of our system; since we do not discuss the generation task here, we presently skip the details of this aspect of the ontology. In the following section we describe the features used for the ontology mapping.

3.1 Feature Set

The interpretation component uses general features that are applicable to any conversation domain. The first set of features we use for ontology mapping are features relating to conversational structure. These are listed and briefly described in Table 1. The $Sprob$ and $Tprob$ features measure how terms cluster between conversation participants and conversation turns. There are simple features measuring sentence length (SLEN, SLEN2) and position (TLOC, CLOC). Pause-style features indicate how much time transpires between the previous turn, the current turn and the subsequent turn (PPAU, SPAU). For email conversations, pause features are based on the timestamps between consecutive emails. Lexical features capture cohesion (CWS) and cosine similarity between the sentence and the conversation (CENT1, CENT2). All structural features are normalized by document length.

Feature ID	Description
MXS	max *Sprob* score
MNS	mean *Sprob* score
SMS	sum of *Sprob* scores
MXT	max *Tprob* score
MNT	mean *Tprob* score
SMT	sum of *Tprob* scores
TLOC	position in turn
CLOC	position in conv.
SLEN	word count, globally normalized
SLEN2	word count, locally normalized
TPOS1	time from beg. of conv. to turn
TPOS2	time from turn to end of conv.
DOM	participant dominance in words
COS1	cos. of conv. splits, w/ *Sprob*
COS2	cos. of conv. splits, w/ *Tprob*
PENT	entro. of conv. up to sentence
SENT	entro. of conv. after the sentence
THISENT	entropy of current sentence
PPAU	time btwn. current and prior turn
SPAU	time btwn. current and next turn
BEGAUTH	is first participant (0/1)
CWS	rough ClueWordScore
CENT1	cos. of sentence & conv., w/ *Sprob*
CENT2	cos. of sentence & conv., w/ *Tprob*

Table 1: Features Key

While these features have been found to work well for generic extractive summarization, we use additional features for capturing the more specific sentence-level phenomena of this research.

- **Character trigrams** We derive all of the character trigrams in the collected corpora and include features indicating the presence or absence of each trigram in a given sentence.
- **Word bigrams** We similarly derive all of the word bigrams in the collected corpora.
- **POS bigrams** We similarly derive all of the POS-tag bigrams in the collected corpora.
- **Word pairs** We consider w_1, w_2 to be a word pair if they occur in the same sentence and w_1 precedes w_2. We derive all of the word pairs in the collected corpora and includes features indicating the presence or absence of each word pair in the given sentence. This is essentially a skip bigram where any amount of intervening material is allowed as long as the words occur in the same sentence.
- **POS pairs** We calculate POS pairs in the same manner as word pairs, above. These are essentially skip bigrams for POS tags.
- **Varying instantiation ngrams** We derive a simplified set of VIN features for these experiments. For each word bigram w_1, w_2, we further represent the bigram as p_1, w_2 and w_1, p_2 so that each pattern consists of a word and a POS tag. We include a feature indicating the presence or absence of each of these varying instantiation bigrams.

After removing features that occur fewer than five times, we end up with 218,957 total features.

4 Transformation - ILP Content Selection

In the previous section we described how we identify sentences that link participants and entities through a variety of sentence-level phenomena. Having populated our ontology with these triples to form a source representation, we now turn to the task of transforming the source representation to a summary representation, identifying the $< participant, relation, entity >$ triples for which we want to generate text. We adapt a method proposed by Xie et al. (2009) for extractive sentence selection. They propose an ILP approach that creates a summary by maximizing a global objective function:

$$maximize\ (1-\lambda) * \sum_i w_i c_i + \lambda * \sum_j u_j s_j \quad (1)$$

$$subject\ to \sum_j l_j s_j < L \quad (2)$$

where w_i is the *tf.idf* score for concept i, u_j is the weight for sentence j using the cosine similarity to the entire document, c_i is a binary variable indicating whether concept i is selected (with the concept represented by a unique weighted n-gram), s_j is a binary variable indicating whether sentence j is selected, l_j is the length of sentence j and L is the desired summary length. The λ term is used to balance concept and sentence weights. This method selects sentences that are weighted strongly and which cover as many important concepts as possible. As described by Gillick et al. (2009), concepts and sentences are tied together by two additional constraints:

$$\sum_j s_j o_{ij} \geq c_i\ \forall_i \quad (3)$$

$$s_j o_{ij} \leq c_i\ \forall_{i,j} \quad (4)$$

where o_{ij} is the occurence of concept i in sentence j. These constraints state that a concept can only be selected if it occurs in a sentence that is selected, and that a sentence can only be selected if all of its concepts have been selected.

We adapt their method in several ways. As mentioned in the previous section, we use weighted noun phrases as our entities instead of n-grams. In our version of Equation 1, w_i is the *tf* score of entity i (the *idf* was already used to identify entities as described previously). More importantly, our sentence weight u_j is the sum of all the posterior probabilities for sentence j derived from the various sentence-level classifiers. In other words, sentences are weighted highly if they correspond to multiple object properties in the ontology. To continue the example from Section 3, the sentence *Let's go with the simple chip* may be selected because it represents both a decision and a positive-subjective opinion, as well as containing the entity *simple chip* which is mentioned frequently in the conversation.

We include constraint 3 but not 4; it is possible for a sentence to be extracted even if not all of its entities are. We know that all the sentences under consideration will contain at least one entity because sentences with no entities would not have been mapped to the ontology in the form of $< participant, relation, entity >$ triples in the first place. To begin with, we set the λ term at 0.75 as we are mostly concerned with identifying important sentences containing multiple links to the ontology. In our case L is 20% of the total document word count.

5 Experimental Setup

In this section we describe our conversation corpora, the statistical classifiers used, and the evaluation metrics employed.

5.1 Corpora

These experiments are conducted on both meeting and email conversations, which we describe in turn.

5.1.1 The AMI Meetings Corpus

For our meeting summarization experiments, we use the *scenario* portion of the AMI corpus (Carletta et al., 2005), where groups of four participants take part in a series of four meetings and play roles within a fictitious company. There are 140 of these meetings in total, including a 20 meeting test set containing multiple human summary annotations per meeting (the others are annotated by a single individual). We report results on both manual and ASR transcripts. The word error rate for the ASR transcripts is 38.9%.

For the *summary annotation*, annotators wrote abstract summaries of each meeting and extracted sentences that best conveyed or supported the information in the abstracts. The human-authored abstracts each contain a general abstract summary and three subsections for "decisions," "actions" and "problems" from the meeting. A many-to-many mapping between transcript sentences and sentences from the human abstract was obtained for each annotator. Approximately 13% of the total transcript sentences are ultimately labeled as extracted sentences. A sentence is considered a decision item if it is linked to the decision portion of the abstract, and action and problem sentences are derived similarly.

For the *subjectivity annotation*, we use annotations of positive-subjective and negative-subjective utterances on a subset of 20 AMI meetings (Wilson, 2008). Such subjective utterances involve the expression of a private state, such as a positive/negative opinion, positive/negative argument, and agreement/disagreement. Of the roughly 20,000 total sentences in the 20 AMI meetings, nearly 4000 are labeled as *positive-subjective* and nearly 1300 as *negative-subjective*.

5.1.2 The BC3 Email Corpus

While our main experiments focus on the AMI meeting corpus, we follow these up with an investigation into applying our abstractive techniques to email data. The BC3 corpus (Ulrich et al., 2008) contains email threads from the World Wide Web Consortium (W3C) mailing list. The threads feature a variety of topics such as web accessibility and planning face-to-face meetings. The annotated portion of the mailing list consists of 40 threads. The threads are annotated in the same manner as the AMI corpus, with three human annotators per thread first authoring abstracts and then linking email thread sentences to the abstract sentences. The corpus also contains speech act annotations. Unlike the AMI corpus, however, there are no annotations for deci-

sions, actions and problems, an issue addressed later.

5.2 Classifiers

For these experiments we use a maximum entropy classifier using the *liblinear* toolkit[1] (Fan et al., 2008). For each of the AMI and BC3 corpora, we perform 10-fold cross-validation on the data. In all experiments we apply a 20% compression rate in terms of the total document word count.

5.3 Evaluation

We evaluate the various classifiers described in Section 3 using the ROC curve and the area under the curve (AUROC), where a baseline AUROC is 0.5 and an ideal classifier approaches 1.

To evaluate the content selection in the transformation stage, we use weighted recall.This evaluation metric is based on the links between extracted sentences and the human gold-standard abstracts, with the underlying motivation being that sentences with more links to the human abstract are generally more informative, as they provide the content on which an effective abstract summary should be built. If M is the number of sentences selected in the transformation step, O is the total number of sentences in the document, and N is the number of annotators, then Weighted Recall is given by

$$recall = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} L(s_i, a_j)}{\sum_{i=1}^{O}\sum_{j=1}^{N} L(s_i, a_j)}$$

where $L(s_i, a_j)$ is the number of links for a sentence s_i according to annotator a_j. We can compare machine performance with human performance in the following way. For each annotator, we rank their sentences from most-linked to least-linked and select the best sentences until we reach the same word count as our selections. We then calculate their weighted recall score by using the other N-1 annotations, and then average over all N annotators to get an average human performance. We report all transformation scores normalized by human performance for that dataset.

6 Results

In this section we present results for our interpretation and transformation components.

[1]http://www.csie.ntu.edu.tw/ cjlin/liblinear/

6.1 Interpretation: Meetings

Figure 1 shows the ROC curves for the sentence-level classifiers applied to manual transcripts. On both manual and ASR transcripts, the classifiers with the largest AUROCs are the action item and general extractive classifiers. Action item sentences can be detected very well with this feature set, with the classifier having an AUROC of 0.92 on manual transcripts and 0.93 on ASR, a result comparable to previous findings of 0.91 and 0.93 (Murray and Renals, 2008) obtained using a speech-specific feature set. General extractive classification is also similar to other state-of-the-art extraction approaches on spoken data using speech features (Zhu and Penn, 2006)[2] with an AUROC of 0.87 on manual and 0.85 on ASR. Decision sentences can also be detected quite well, with AUROCs of 0.81 and 0.77. Positive-subjective, negative-subjective and problem sentences are the most difficult to detect, but the classifiers still give credible performance with AUROCs of approximately 0.76 for manual and 0.70-0.72 for ASR.

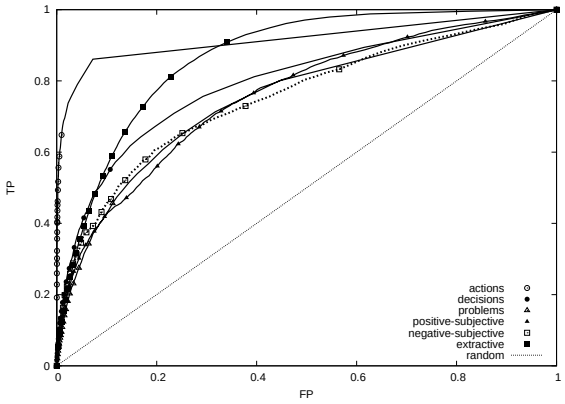

Figure 1: ROC Curves for Ontology Mapping Classifiers (Manual Transcripts)

6.2 Transformation: Meetings

In this section we present the weighted recall scores for the sentences selected using the ILP method described in Section 4. Remember, weighted recall measures how useful these sentences would be in generating sentences for an abstract summary. We also assess the performance of three baseline summarizers operating at the same compression level.

[2]Based on visual inspection of their reported best ROC curve

The simplest baseline (GREEDY) selects sentences by ranking the posterior probabilites output by the general extractive classifier. The second baseline (CLASS COMBO) averages the posterior probabilites output by *all* the classifiers and ranks sentences from best to worst. The third baseline (RETRAIN) uses the posterior probability outputs of all the classifiers (except for the extractive classifier) as new feature inputs for the general extractive classifier.

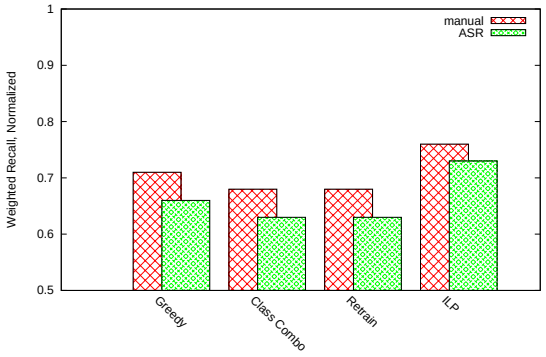

Figure 2: Weighted Recall Scores for AMI Meetings

Figure 2 shows the weighted recall scores, normalized by human performance, for all approaches on both manual and ASR transcripts. On manual transcripts, the ILP approach (0.76) is better than GREEDY (0.71) with a marginally significant difference (p=0.07) and is significantly better than CLASS COMBO and RETRAIN (both 0.68) according to t-test ($p < 0.05$) . For ASR transcripts, the ILP approach is significantly better than all other approaches ($p < 0.05$). Xie et al. (2009) reported ROUGE-1 F-measures on a different meeting corpus, and our ROUGE-1 scores are in the same range of 0.64-0.69 (they used 18% compression ratio).

6.3 Interpretation: Emails

We applied the same summarization method to the 40 BC3 email threads, with contrasting results. Because the BC3 corpus does not currently contain annotations for decisions, actions and problems, we simply ran the AMI-trained models over the data for those three phenomena. We can assess the performance of the extractive, positive-subjective and negative-subjective classifiers by examining the ROC curves displayed in Figure 3. Both the general extractive and negative-subjective classifiers have AUROCs of around 0.75. The positive-subjective classifier initially has the worst performance with an AUROC of 0.66, but we found that positive-subjective performance increased dramatically to an AUROC of 0.77 when we used only conversational features and not word bigrams, character trigrams or POS tags.

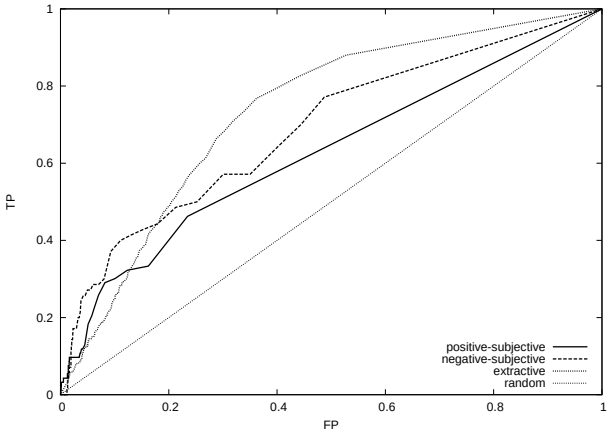

Figure 3: ROC Curves for Ontology Mapping Classifiers (BC3 Corpus)

6.4 Transformation: Emails

If we examine the weighted recall scores in Figure 4 we see that the ILP approach is worse than the greedy summarizers on the BC3 dataset. However, the differences are not significant between ILP and COMBO CLASS (p=0.15) and only marginally significant compared with RETRAIN and GREEDY (both p=0.08). The performance of the ILP approach varies greatly across email threads. The top 15 threads (out of 40) yield ILP weighted recall scores that are on par with human performance, while the worst 15 are half that.

6.4.1 Email Corpus Analysis

Due to the large discrepancy in performance on BC3 emails, we conducted additional experiments for error analysis. We first explored whether we could build a classifier that could discriminate the best 15 emails from the worst 15 emails in terms of weighted recall scores with the ILP approach, to determine whether there are certain features that correlate with good performance. Using the same fea-

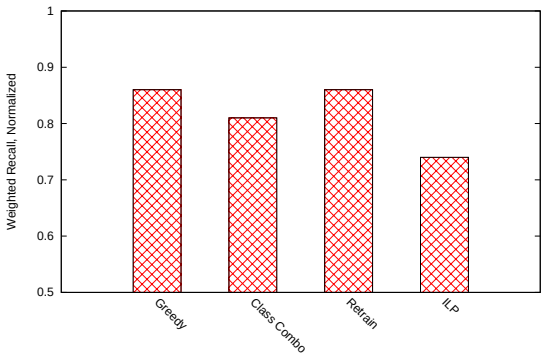

Figure 4: Weighted Recall Scores for BC3 Threads

tures described in Section 3.1, we built a logistic regression classifier on the two classes and found that they can be discriminated quite well (80% accuracy on an approximately balanced dataset) and that the conversation structure features are the most useful for discerning them. Table 2 shows the weighted recall scores and several conversation features that were weighted most highly by the logistic regression model. In particular, we found that the email threads that yielded good performance tended to feature more active participants (# Participants), were not dominated by a single individual (BEGAUTH), and featured a higher number of turns (# Turns) that followed each other in quick succession without long pauses (PPAU, pause as percentage of conversation length). In other words, these emails were structured more similarly to meetings. Note that since we normalize weighted recall by human performance, it is possible to have a weighted recall score higher than 1. On the 15 best threads, our system achieves human-level performance. Because we used AMI-trained models for detecting decisions, actions and problems in the BC3 data, it is not surprising that performance was better on those emails structured similarly to meetings. All of this indicates that there are many different types of emails and that we will have to focus on improving performance on emails that differ markedly in structure.

7 Conclusion

We have presented two components of an abstractive conversation summarization system. The *interpretation* component is used to populate a simple conversation ontology where conversation participants and entities are linked by object properties such as decisions, actions and subjective opinions. For this step we show that highly accurate classifiers can be built using a large set of features not specific to any conversation modality.

Metric	Worst 15	Best 15
Weighted Recall	0.49	1.05
# Turns	6.27	6.73
# Participants	4.67	5.4
PPAU	0.18	0.12
BEGAUTH	0.31	0.18

Table 2: Selected Email Features, Averaged

In the *transformation* step, a summary is created by maximizing a function relating sentence weights and entity weights, with the sentence weights determined by the sentence-ontology mapping. Our evaluation shows that the sentences we select are highly informative to generate abstract summaries, and that our content selection method outperforms several greedy selection approaches. The system described thus far may appear extractive in nature, as the transformation step is identifying informative sentences in the conversation. However, these selected sentences correspond to $< participant, relation, entity >$ triples in the ontology, for which we can subsequently generate novel text by creating linguistic annotations of the conversation ontology (Galanis and Androutsopolous, 2007). Even without the generation step, the approach described above allows us to create structured extracts by grouping sentences according to specific phenomena such as action items and decisions. The knowledge represented by the ontology enables us to significantly improve sentence selection according to intrinsic measures and to generate structured output that we hypothesize will be more useful to an end user compared with a generic unstructured extract.

Future work will focus on the generation component and on applying the summarization system to conversations in other modalities such as blogs and instant messages. Based on the email error analysis, we plan to pursue domain adaptation techniques to improve performance on different types of emails.

References

R. Barzilay and K. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

G. Carenini, R. Ng, and X. Zhou. 2007. Summarizing email conversations with clue words. In *Proc. of ACM WWW 07, Banff, Canada*.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI meeting corpus: A preannouncement. In *Proc. of MLMI 2005, Edinburgh, UK*, pages 28–39.

K. Church and W. Gale. 1995. Inverse document frequency IDF: A measure of deviation from poisson. In *Proc. of the Third Workshop on Very Large Corpora*, pages 121–130.

J. Clarke and M. Lapata. 2006. Constraint-based sentence compression: An integer programming approach. In *Proc. of COLING/ACL 2006*, pages 144–151.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

D. Galanis and I. Androutsopolous. 2007. Generating multilingual descriptions from linguistically annotated owl ontologies: the naturalowl system. In *Proc. of ENLG 2007, Schloss Dagstuhl, Germany*.

M. Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pages 364–372.

D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tür. 2009. A global optimization framework for meeting summarization. In *Proc. of ICASSP 2009, Taipei, Taiwan*.

S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky. 2007. Resolving *"You"* in multi-party dialog. In *Proc. of SIGdial 2007, Antwerp, Belgium*.

L. He, E. Sanocki, A. Gupta, and J. Grudin. 1999. Auto-summarization of audio-video presentations. In *Proc. of ACM MULTIMEDIA '99, Orlando, FL, USA*, pages 489–498.

P-Y. Hsueh, J. Kilgour, J. Carletta, J. Moore, and S. Renals. 2007. Automatic decision detection in meeting speech. In *Proc. of MLMI 2007, Brno, Czech Republic*.

K. Spärck Jones. 1999. Automatic summarizing: Factors and directions. In I. Mani and M. Maybury, editors, *Advances in Automatic Text Summarization*, pages 1–12. MITP.

T. Kleinbauer, S. Becker, and T. Becker. 2007. Combining multiple information layers for the automatic generation of indicative meeting abstracts. In *Proc. of ENLG 2007, Dagstuhl, Germany*.

K. Knight and D. Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proc. of AAAI 2000, Austin, Texas, USA*, pages 703–710.

J. Kupiec, J. Pederson, and F. Chen. 1995. A trainable document summarizer. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA*, pages 68–73.

K. McKeown, J. Hirschberg, M. Galley, and S. Maskey. 2005. From text to speech summarization. In *Proc. of ICASSP 2005, Philadelphia, USA*, pages 997–1000.

C. Muller. 2007. Resolving *It*, *This* and *That* in unrestricted multi-party dialog. In *Proc. of ACL 2007, Prague, Czech Republic*.

G. Murray and S. Renals. 2008. Detecting action items in meetings. In *Proc. of MLMI 2008, Utrecht, the Netherlands*.

G. Murray, T. Kleinbauer, P. Poller, S. Renals, T. Becker, and J. Kilgour. 2008. Extrinsic summarization evaluation: A decision audit task. In *Proc. of MLMI 2008, Utrecht, the Netherlands*.

M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium*.

S. Raaijmakers, K. Truong, and T. Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proc. of EMNLP 2008, Honolulu, HI, USA*.

O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen. 2004. Summarizing email threads. In *Proc. of HLT-NAACL 2004, Boston, USA*.

J. Ulrich, G. Murray, and G. Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *Proc. of AAAI EMAIL-2008 Workshop, Chicago, USA*.

T. Wilson. 2008. Annotating subjective content in meetings. In *Proc. of LREC 2008, Marrakech, Morocco*.

S. Xie, B. Favre, D. Hakkani-Tür, and Y. Liu. 2009. Leveraging sentence weights in a concept-based optimization framework for extractive meeting summarization. In *Proc. of Interspeech 2009, Brighton, England*.

L. Zhou and E. Hovy. 2005. Digesting virtual "geek" culture: The summarization of technical internet relay chats. In *Proc. of ACL 2005, Ann Arbor, MI, USA*.

X. Zhu and G. Penn. 2006. Summarization of spontaneous conversations. In *Proc. of Interspeech 2006, Pittsburgh, USA*, pages 1531–1534.