Automatic Summarization of Conversational Multi-party Speech

Ph.D. Thesis Proposal

Michel Galley
Department of Computer Science
Columbia University
galley@cs.columbia.edu

Abstract

This proposal addresses the problem of automatically summarizing conversational speech, in particular meeting recordings. The problem is divided into two main steps: *utterance selection*, the task of identifying a set of utterances representative of the important elements of a meeting, and *utterance revision*, the task of creating fluent and concise utterances from the ones produced by a speech recognizer.

I propose a discourse-based approach to utterance selection that incorporates two processing stages: the first stage is to segment the meeting transcription by topic, a process that provides a high-level structure to the summary to be generated. The second stage analyzes each topical segment and attempts to predict the communicative goal (dialog act) of each utterance in order to determine, given a pragmatic context defined by preceding and succeeding dialog acts, whether the utterance should be included in the summary or not. The second stage is realized using dynamic Bayesian networks, a computational framework that combines here surface features known to be good predictors in the summarization task and inter-sentential discourse dependencies. This enables selected utterances to fit their summaries in coherent discourse situations.

The proposed utterance revision system uses various knowledge sources and syntactic transformation rules automatically acquired from document-abstract pairs. It transforms recognized utterances into more readable and concise forms, by removing uninformative, syntactically optional constituents, and disfluencies. Knowledge sources that are used to score revision hypotheses include: n-gram and syntax-based language models, a constituent-removal probability model, a model of prosodic markedness, and various models of constituent importance derived from $tf \cdot idf$ scores. The weights of all models are discriminatively trained to optimize summary quality. Models are finally combined to select a globally optimal analysis, which can be efficiently selected using a CKY-style decoding algorithm.

Contents

List of Figures

List of Tables

1 Introduction

Document summarization has proved to be a desirable component in many information management systems, complementing core information retrieval and browsing functionalities. The use of document summarization techniques is especially important with speech documents such as meeting transcriptions, since they are particularly difficult to navigate. As opposed to their written counterparts, spoken documents lack structural elements like title, headers, and paragraph that can ease the task of the information seeker. Document summarization can reduce the overhead of navigating large collections of speech data. For example, summarization techniques can be used to extract or highlight relevant passages in the full transcription. Alternatively, similar techniques might be used to generate short text substitutes, or abstracts, that capture the "aboutness" of the meeting, while discarding their disfluencies and unimportant constituents.

In this proposal, I address the problem of creating abstractive summaries of meetings, a problem that has to date only received limited attention.[1] Abstractive summaries differ from extractive summaries, in that they are not simple concatenations of input utterances, but alter the extracted material in order to produce a fluent and concise summary.

Speech summarization in general, and meeting summarization in particular, faces various issues that differ from text summarization:

- Meetings are notoriously hard to recognize automatically: the word error rate (WER) of a state-of-the-art automatic speech recognition (ASR) system [Mirghafori *et al.*, 2004] on standard meeting test data [NIST, 2004b] is 34.8%. Furthermore, disfluencies, ungrammaticalities, and other speech errors are frequent in spontaneous speech, which make it particularly difficult to directly re-use ASR transcriptions to create summaries.
- As opposed to text, there are no trivially recognizable elements indicative of overall structure, such as headlines, sections, paragraphs, or punctuation that may help utterance selection. Traditional text summarization techniques are inadequate in this case, since they markedly rely on this kind of information.
- In conversational speech, there is generally only limited evidence of pre-planned speech: while interaction generally conforms to regularities defined by turn-taking rules, common conversational patterns such as question-answer pairs, and politeness [Levinson, 1983], conversation unfolds in complex interplays that generally reflect more the personality of those involved in the conversation (e.g. whether they will agree or cause prolonged discussion, or whether they will receive a particular comment well) than any initial plan any participant may have [Hughes, 1996, ch. 1]. This apparent lack of coherence make it difficult to identify which speech contributions are important, since coherence and strategies used by human abstractors are closely tied (see, e.g., discourse-based summarization approaches in [Mani and Maybury, 1999]).

Speech summarization presents, however, some opportunities not found with text (see [McKeown *et al.*, 2005] for a comparison between text and speech summarization). Additional information can be extracted from the speech signal, in particular pitch contour, pitch range, energy, pausal duration, speech rate, and frequency of speaker overlap (competitive speech), some of which were

[1] Zechner and Waibel [2000] perform summarization of meetings and other speech genres with a technique that is essentially extractive – the speech extract is only edited to remove speech disfluencies

found useful in various speech summarization tasks [Inoue *et al.*, 2004; Maskey and Hirschberg, 2005]. One goal of the proposed work is to study the acoustic correlate of perceived salience in the context of meeting summarization.

Given the problems pertaining to extractive techniques listed above, it seems reasonable to divide the meeting summarization task into two main steps: *utterance selection*, i.e. identifying a set of salient utterances that is a practical substitute to the entire meeting transcription; *utterance revision*, i.e. correcting the various non-fluencies typical of conversational speech, and operating below the sentence level to further remove unimportant lexical material.

I propose an approach to **utterance selection** that incorporates two processing stages: the first stage is to divide the meeting transcription into a linear sequence of topical segments, a structure supposed to mirror the high-level organization of each meeting. This accounts for the fact that meetings under investigation have pre-set agendas, and that, even though topical segments may seem chaotic or underpinned by complex interplays, their segmentation generally seem to be well-defined (human labelers reached marked agreement in identifying them[Galley *et al.*, 2003]). The second stage builds a quite "surfacy" discourse representation of each segment by assigning to each utterance a dialog act (DA) label that characterizes its communicative goal, a representation that was found to strongly correlate with summaries built by humans. This representation is concretely realized as a *dynamic Bayesian network* (DBN) structure incorporating two Markov processes – DA labels and summary labels (SUM and NON-SUM) – where the probability of including any given utterance can effectively be conditioned on a pragmatic context of previous DAs.

The main research contribution of the proposed work in utterance selection lies in its use of a probabilistic model of interpersonal interaction to automatically determine the structure of the DBN, i.e. probabilistically determine at each point in time the set of parent nodes of each dynamic variable. The interaction model currently only uses speaker-addressee relations to shape the structure of the DBN, but it is conceptually quite simple to extend it to deal with other types of relations (possibly genre or domain specific), e.g. employer-employee and advisor-student. The "switching-parent" DBN structure, the interpersonal discourse model have characteristics that are quite novel to summarization. Preliminary experiments with statistical tests confirmed that the dependencies introduced by the proposed DBNs are grounded by empirical evidence, and classification results with different "oracle" classifiers (i.e. using true contextual labels) indicate positive prospects, and levels of performance substantially close to humans.

I propose a syntax-driven **utterance revision** system that is aimed at condensing utterances by removing non-fluencies typical to spontaneous speech, as well as semantically empty phrases (e.g. "you know"), unimportant, and grammatically optional constituents. Its CKY decoder exploits syntactic transformation rules automatically acquired from transcription-abstract parallel texts (e.g. adjunct dropping rules) with a syntax-based language model to promote condensation hypotheses that are both short and grammatical. I also propose to examine the use of different parameter estimation techniques to incorporate an arbitrary number of additional knowledge sources, such as semantic role labels, models of lexical salience (e.g. $tf \cdot idf$), and prosody to exploit any existing acoustic correlates of words found in abstracts, e.g. prosodically stressed words.

The proposed research in utterance revision encompasses the following contributions which are novel in revision-base summarization: 1) a representation of revision rules that is more general than the kind of paired CFG grammars which have commonly been used in previous sentence condensation systems. The extended domain of locality pertaining of these rules allow us to model the transformation between *any* tree pair as a sequence of revision rules. 2) The use of an algorithm

to find the minimum tree edit distance alignment between the pair, which we use in turn to define a decomposition of the pair into a sequence of revision operations. 3) a head-driven lexicalized parametrization for revision rules that accounts for the fact that certain words (e.g. "not") are less likely to be deleted than other words with same grammatical role (e.g. "somehow"). 4) A discriminative training framework in which the parameter of different models can be optimized for the given task.

The main issues still to be addressed in this research are:

- Determine what form of discourse representation (e.g., dialog acts) is the most effective for utterance selection. While standard dialog act sets such as DAMSL [Allen and Core, 1997] exist, task-specific sets may be more appropriate for the task at hand.
- Determine what kind of dependencies need to be encoded into dynamic Bayesian networks in utterance selection.
- Determine how to effectively weight the different models used in utterance revision, and investigate the contribution of each knowledge source.
- Evaluate meeting summaries, both automatically, during training and development, and manually, in a final evaluation.

This proposal is divided as follows: Section 2 describes the meeting data used to train and test a meeting summarizer, as well as other corpora that may be used to train the various models used by the summarizer. Section 3 describes two topic segmentation algorithms, and their use in the summarization system. Section 4 and 5 respectively describe the utterance selection and revision components of the summarizer.

2 Corpus

The work proposed for my thesis is primarily meant to be applied to the ICSI corpus [Janin *et al.*, 2003], a corpus of naturally-occurring meetings, i.e. meetings that would have taken place anyway. Their style is quite informal, and topics were primarily concerned with speech, natural language, or artificial intelligence, and networking research. The corpus contains 75 meetings, which are 60 minutes long on average, and involve a number of participants ranging from 3 to 10 (6 on average). The total number of unique speakers is 60, including 26 non-native English speakers.

The speech of each participant was recorded on a separate channel with a close-talking microphone (generally head mounted). Human word-level orthographic transcriptions are available for all meetings, and incorporate punctuation and the annotation of various speech and non-speech events (e.g. laughs and door slams). Each dialog turn was labeled with speaker ID, start time, and end time. Furthermore, all meetings were automatically recognized using the ICSI-SRI-UW speech recognition system, a conversational telephone speech (CTS) recognizer with language and acoustic models adapted to the meeting domain.[2] Recognition word-lattices with separate acoustic and language model scores are also available for the 75 meetings.

Furthermore, human annotation of topic segmentation [Galley *et al.*, 2003; Carletta *et al.*, 2003], DA labels, and adjacency pairs (AP), i.e. paired utterances such as question-answer, and

[2] [Mirghafori *et al.*, 2004] reports a WER of 24.2% on ICSI test data, while the system's WER on the entire NIST RT-04 evaluation data is 34.8%.

apology-downplay, [Shriberg *et al.*, 2004] is available for all 75 meetings. An ongoing labeling effort at University of Edinburgh aims to create both extractive and abstractive summaries [Carletta *et al.*, 2003]. For extractive summaries, human judges were asked for all 75 meetings to select transcription utterances to include in summaries (compression factors range from 90 to 95%, though no strict limit was imposed). Nine meetings were summarized by more than two, and agreement level was quite low ($\kappa = .315$). The data currently contains 61 abstractive summaries, which are made of four sections: "abstract", "decisions", "problems", and "progress and achievements". They have a quite high abstraction level, and seldom re-use phrases of the transcriptions.

The increased level of interest in meeting processing and recent meeting recognition task evaluations [NIST, 2004b] favored the emergence of additional meeting corpora. The CMU/ILS corpus [Waibel *et al.*, 2001] contains 18 transcribed meetings that were either naturally occurring (research group meetings) or artificial (game playing tasks, military strategic exercises, discussions of general topics including news [Burger *et al.*, 2002]). The NIST pilot corpus [NIST, 2004a] consists of 19 transcribed meetings, a mix of real meetings (staff regular meetings, technical presentations, planning events) and scenario-driven meetings (game playing, focus groups discussions about general subjects such as news events and movies). The AMI corpus contains meetings recorded mostly at IDIAP [McCowan *et al.*, 2003] representing about 100 hours of meeting time. Some are naturally occurring (research meetings), others are elicited, e.g. task-oriented meetings where participants play different roles in design teams (see [Carletta *et al.*, 2005]), and informal discussions about freely chosen subjects (e.g. discussions about recent travels, last books and papers read, and movies). The corpus is not yet officially released, and currently not all meetings are transcribed.[3] Rich transcription (RT) NIST evaluations [NIST, 2004b] rely on the ICSI, CMU, and NIST data, and define standard test sets for these corpora.

3 Topic segmentation

Meetings in the ICSI corpus are topically very diverse, and it is admittedly desirable to segment them into topically cohesive units before proceeding to the summarization step, and to produce topically organized summaries. Indeed, the alternative approach consisting of producing undivided summaries does not seem particularly appealing if we consider that reference extractive summaries extracts produced by my preliminary system (see Section 4) remain relatively long, even at high compression levels (a 95%-compression level results in summaries of 84 utterances on average.)

Another reason for performing segmentation by topic lies in the statistical correlation between topic segments and summary labels: as Figure 1 suggests, an utterance appearing at the beginning of a topical segment has higher chances of being included in a summary than utterances appearing elsewhere in the segment. The correlation between topic segmentation and summarization is significant, since the null hypothesis that utterances on the first 10% of each segment and those of the remaining 90% are equally likely to be included in a summary is clearly rejected.[4] This observation is consistent with similar findings in text summarization that established that positional

[3]The current release is available from: `http://mmm.idiap.ch`

[4]A χ^2 test demonstrates that the differences between the two observed frequencies (1-10% vs. 11-100%) measured against the dichotomous classification (summary vs. non-summary) is highly significant ($\chi^2 = 249.1$, $df = 1$, $p < .001$). The same test with between the 10 observed frequencies (1-10%, 11-20%, etc) and the same dichotomous classification reaches the same conclusion ($\chi^2 = 358.6$, $df = 9$, $p < .001$).

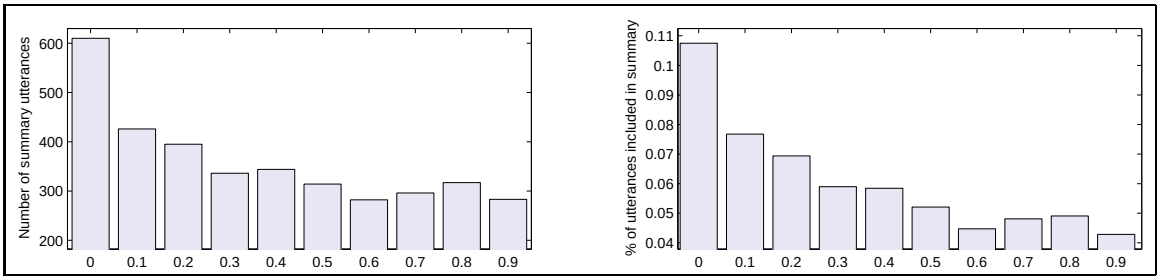

Figure 1: Histogram of summary sentences tabulated by relative position (x-axis) with respect to the reference topic segmentation (first bin corresponds to the first 10% of the segment, etc.) The y-axis of the left plot displays raw counts of in-summary utterances, and, in the right plot, it represents percentages of utterances that are included in the summaries.

features are generally strongly correlated with what is included in a summary, and that sentences of the lead section of a document are generally good candidates for inclusion.

3.1 Related work

Previous work in topic segmentation is particularly extensive for both speech and text (see [Reynar, 1998] for an overview, and [Galley *et al.*, 2003] for more recent techniques). Approaches have predominantly used off-the-shelf learning algorithms, particularly decision trees and Bayesian models, to exploit a variety of local predictors that may indicate topic boundaries, such as acoustic and durational features (pauses, F0, speaking rate), and various lexical features such as cue words and referential noun phrases. These approaches are sometimes referred to as feature-based algorithms.

Other approaches particularly prominent in text segmentation exploit the fact that topic segments tend to be lexically cohesive, i.e. the same terms tend to be repeated many times within the same topic. These approaches capitalize on lexical distributional information and have used various criteria to identify topic boundaries, e.g. cosine similarity between word-vectors built from two adjacent windows (TextTiling) [Hearst, 1994], cosine similarity and Kullback-Leibler (KL) distance applied to dotplots or self-similarity matrices [Reynar, 1998; Choi, 2000; Renals and Ellis, 2003]. These techniques essentially differ from feature-based algorithms by their use of statistics gathered from the entire text or recording to make segmental decisions that are deemed globally optimal.

The work presented in [Galley *et al.*, 2003] investigates some approaches to incorporate lexical statistics in a feature-based segmenter (decision trees) to make it sensitive to lexical cohesion. In this work, we designed a feature utilizing two adjacent windows that quantify the degree of lexical cohesion at the boundary between windows by computing cosine similarity between the two.[5] The lexical cohesion feature is incorporated along with other features, some of which are specific to multi-party speech and novel.

[5] As opposed to Hearst [1994], who computed cosine similarity between pairs of word vectors, the similarity here is computed based on chains of repeated terms: if a strong chain repetition overlaps the two windows, cosine similarity will tend to be higher; conversely, a chain ending in the left window or starting in the right one will decrease the similarity score. Details are left in the paper [Galley *et al.*, 2003].

3.2 Sequential binary classification

Topic segmentation is cast here as a binary classification problem where each potential boundary b_i (e.g. utterance break) in a sequence $\mathbf{b} = < b_1, ..., n_T >$ is either considered a topic boundary or not ($b_i \in \{0, 1\}$).

I applied two classical supervised learning algorithms (C4.5 and support vector machines) to build classifiers that use lexical, durational, and distributional (e.g. lexical cohesion) features to determine if an utterance break corresponds to a topic boundary; those features are described in Section 3.3, and results are presented in Section 3.5.

For every potential boundary b_i, the classifier assumes that only a fixed-size neighborhood of observations $o_{i,t}$ is relevant in order to determine the class label of b_t, e.g. the classifier will count the number of occurrences of a given cue word within a certain range of the potential boundary. It is generally unclear what is the optimal neighborhood or window size and how features should be analyzed within the window (e.g. should we take the mean, minimum, or maximum value?). Windows of various sizes can lead to different levels of prediction, and in some cases, it might be more appropriate to only extract features preceding or following b_i. Too small windows may overlook longer-range interactions between class labels and observations. Too large windows may not exploit well observations that are only highly predictive within short ranges, and cause consecutive class variables b_i to be close in feature space, making it more difficult to find good decision boundaries to separate positive instances from negative ones.

I avoided making arbitrary choices regarding the size of feature-extraction window. For each feature, I greedily searched for the optimal window size, optimizing the error metric described in Section 3.4. For each feature I also determined if only the left context, right context, or both were useful. Search results for features presented in Section 3.3 are described in more details in [Galley *et al.*, 2003].

3.3 Features for topic segmentation

The topic segmenter presented in [Galley *et al.*, 2003] exploited the following features:

Cue phrases: previous work on segmentation has found that discourse particles like *now*, *well* provide valuable information about the structure of texts [Grosz and Sidner, 1986; Hirschberg and Litman, 1994; Passonneau and Litman, 1997]. I analyzed the correlation between words in the meeting corpus and labeled topic boundaries, and automatically extracted utterance-initial cue phrases[6] that are statistically correlated with boundaries. For every word in the meeting corpus, I counted the number of occurrences near any topic boundary (within 10 seconds), and its number of appearances overall. Then, I performed χ^2 significance tests (e.g. figure 1 for *okay*) under the null hypothesis that no correlation exists. I selected terms whose χ^2 value rejected the hypothesis under a .01-level confidence (the rejection cutoff is $\chi^2_{1;.01} = 6.635$). Finally, induced cue phrases whose usage has never been described in other work were removed (marked with $*$ in Table 2). Indeed, there is a risk that the automatically derived list of cue phrases could be too specific to the word usage in these meetings.

Silences: previous work has found that major shifts in topic typically show longer silences [Passonneau and Litman, 1993; Hirschberg and Nakatani, 1996]. I investigated the presence of

[6] As in [Litman and Passonneau, 1995], I restricted myself to the first lexical item of any utterance, plus the second one if the first item is also a cue word.

	$t \leq 10$	$t > 10$
okay	64	740
Other	657	25896

Table 1: "okay" ($\chi^2 = 89.11, df = 1, p < 0.01$).

okay	93.05	but	13.57
shall $*$	27.34	so	11.65
anyway	23.95	and	10.99
we're $*$	17.67	should $*$	10.21
alright	16.09	good $*$	7.70
let's $*$	14.54		

Table 2: List of automatically selected cue phrases.

silences in meetings and their correlation with topic boundaries, and found it helpful to make a distinction between pauses and gaps. According to [Levinson, 1983], a pause is a silence that is attributable to a given party, for example in the middle of an adjacency pair, or when a speaker pauses in the middle of her speech. Gaps are silences not attributable to any party, and last until a speaker takes the initiative of continuing the discussion. As an approximation of this distinction, I classified a silence that follows a question or in the middle of somebody's speech as a pause, and any other silences as a gap. While the correlation between long silences and discourse boundaries seem to be less pervasive in meetings than in other speech corpora, I noticed that some topic boundaries are preceded (within some window) by numerous gaps. However, I found little correlation between pauses and topic boundaries.

Overlaps: I also analyzed the distribution of overlapping speech by counting the average overlap rate within some window. I noticed that, many times, the beginning of segments are characterized by having little overlapping speech.

Speaker change: As Figure 2 suggests, there is a clear correlation between topic boundaries and speech activity (i.e. how long each participant gets to speak within a certain time frame). It is clear that the contribution of individual speakers to the discussion can greatly change from one topic segment to the next. I try to capture significant changes in speakership by measuring the dissimilarity between two consecutive analysis windows, i.e. for each potential boundary, I compute for each speaker i the amount of speech (l_i) and after (r_i) the potential boundary (the size of the two windows is fixed and was optimized using cross validation). The two distributions are normalized to form two probability distributions l and r, and significant changes of speakership are detected by computing their Jensen-Shannon divergence (see e.g. [Lin, 1991]).[7]

Lexical cohesion: I also incorporated the lexical cohesion function referred to previously. Note that I use both the posterior estimate computed by the lexical cohesion algorithm[8] and the raw lexical cohesion cosine similarity value as features of the classifier.

3.4 A probabilistic error metric

Precision and recall are popular statistics for assessing the quality of classification algorithms. These statistics, however, are arguably not well suited for segmentation tasks [Beeferman *et al.*, 1999; Pevzner and Hearst, 2002]. A good segmentation metric would ideally give a relatively high

[7]The Jensen-Shannon divergence is defined as $JS(l, r) = \frac{1}{2}[D(l\|m) + D(r\|m)]$, where $KL(l\|r)$ is the KL-divergence between the two distributions and $m = \frac{1}{2}(l + r)$.

[8]Hearst in [Hearst, 1994] describes a simple technique of analyzing a lexical cohesion cosine similarity plot and computes the relative offset between a given local minimum and the two adjacent local maxima to provide some kind of posterior 'probability' that rewards deep 'valleys' in the plot.

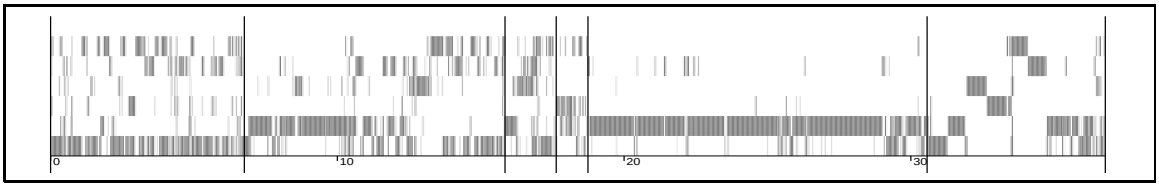

Figure 2: Multi-speaker speech activity across time. Each row represent the speech activity of one speaker, utterance of words being represented as black. Vertical lines represent topic shifts.

score to a segmentation algorithm that systematically comes close real boundaries – say always off by one sentence, and penalize more algorithms that are always far off (precision and recall fail to make this distinction, since scores are zero in all cases).

An ingenious method suggested by John Lafferty [Beeferman *et al.*, 1997], and subsequently used in TDT evaluations [Allan *et al.*, 1998], is based on a probabilistic error metric (lower score is better) that avoids dealing with boundaries explicitly. In our case, the metric estimates the probability that two meeting times selected at random are incorrectly classified as to whether they belong to the same topical segment. Assuming the $\sigma_{\mathtt{s}}(i,j)$ function returns 1 if and only if meeting times i and j are in the same segment according to segmentation $\mathtt{s}$, and that operator $\overline{\oplus}$ represents the boolean function XNOR (both or neither true), the error metric is defined as follows:

$$P_D(\mathtt{ref},\mathtt{hyp}) = \iint\limits_{1\leq i\leq j\leq n} D(i,j)\left(\sigma_{\mathtt{ref}}(i,j)\,\overline{\oplus}\,\sigma_{\mathtt{hyp}}(i,j)\right) \tag{1}$$

The function D is a distance probability distribution, which is commonly set to $D_k = \delta(i+k,j)$, i.e. all the probability mass is set at fixed distance k. The metric has some interesting properties. If for example the distance k is taken to be half of the average reference segment length, the metric assigns $P_k \equiv P_{D_k} \approx .5$ to any algorithm that hypothesizes boundaries at every time or no boundary at all. Similarly, the average P_k will tend to $.5$ if the algorithm assigns boundaries totally randomly. If `ref` and `hyp` are the same, P_k is 0 regardless how D is chosen. The next subsection will present an empirical evaluation of my segmenter using P_k as defined here.

3.5 Results

In [Galley *et al.*, 2003], I used C4.5 [Quinlan, 1993] to build decision trees. I performed double 25-fold cross-validation for evaluating the induced probabilistic classifier, computing the P_k average with the meeting that is held-out at each step (this was a double cross-validation, since for each of the 25 meetings that was held out, I performed a 24-fold cross validation to optimize window sizes and other meta parameters.) I compared the topic segmenter described in this section against a state-of-the-art text segmentation algorithm that is publicly available [Utiyama and Isahara, 2001], and that capitalizes on lexical cohesion only. The quantitative results summarized in Table 3, which correspond to the average performance on the held-out sets, show that the integration of conversational features with the text-based segmenter outperforms the text-based segmenter.

	P_k
C4.5 classifier	23.00%
C4.5 classifier (lexical cohesion only)	31.91%
U00	37.39%
p-value	< 0.001

Table 3: Performance of the feature-based segmenter with decision trees (C4.5).

3.6 Discussion and future work

The topic segmentation module as presented in this section is essentially completed work. I will extend the work and use a larger feature set, particularly add acoustic and prosodic features, to enable comparisons with other speech domains for which all these features (discussed or prosodic/acoustic) have been extensively studied in other speech domains [Tür *et al.*, 2001].

If time permits, I will additionally address the following problem: the main problem with segmentation algorithms trained as binary classifiers as the one presented above is that their objective functions are typically designed to minimize either directly or indirectly classification errors, and as such do not distinguish false alarms that are only one potential boundary away from a true topic boundary from those that are far distant from any other true boundary, and as such, sometimes perform quite poorly with evaluation metrics that are sensitive to near misses, such as the P_k metric. Another main issue with training such segmentation algorithms is that the class distribution is particularly skewed – in the case of the meeting data, positive instances represent about .8% of all training instances, which is problematic for most learning algorithms.

Both problems mentioned above are typically solved by ad-hoc techniques that are theoretically not appealing. Since near-boundary negative instances (e.g. within 10 seconds of a topic shift) are close to positive samples in the feature space, it is often quite effective to remove these near-boundary negative instances; this typically makes it easier to separate the two classes. Furthermore, positive instances need to be up-sampled in most cases to avoid training a classifier that always predicts non-boundaries (this was particularly needed with C4.5).

I propose to study more closely the learning problem as it is formulated for the segmentation problem, and propose training methods that are more sensitive to near misses, and that address the class skewness problem. One particular way of solving this problem is to try to minimize P_k, and various gradient-free optimization algorithms may be used [Wright, 1996], e.g. "downhill" simplex [Nelder and Mead, 1965] or Powell's direction set method [Powell, 1964].

4 Utterance selection

The purpose of content selection is to reduce the input speech transcription to a set of utterances that capture what is important in the source transcription. This will be the core summarization component, and as such it needs to be configurable for any supplied compression ratio (typically 90-95%). The next component in the pipeline (Section 5) is primarily aimed at rendering the output summary into fluent text, and thus will not significantly further reduce the selected text.

Previous work in sentence and utterance selection ranges from methods exploiting surface features and statistics (e.g. position within document, word frequency, and $tf \cdot idf$) [Kupiec *et al.*,

1995] to more elaborate techniques – generally domain specific – involving attempts to devise the meaning of texts [Radev and McKeown, 1998; Hahn and Reimer, 1999].[9] Open-domain semantic interpretation is still beyond the state-of-the-art, so I will rule out techniques involving any semantic analysis; not only does the ICSI meeting corpus range over a quite broad range of topics, but it is also desirable to devise summarization methods easily adaptable to other domains, e.g. business meetings. Alternatively, so-called discourse-based approaches exploit cohesion and coherence properties of discourse to build intermediate representations, e.g. rhetorical tree structures and lexical chains, that enable more sophisticated content reduction operations [Spärck Jones, 1993; Barzilay and Elhadad, 1997; Marcu, 1997; Boguraev and Kennedy, 1997].

I propose the use of a discourse-based approach to utterance selection grounded on the theory of speech acts [Searle, 1969], where utterances are assigned dialog act (DA) labels that characterize their primary communicative goals, such as STATEMENT, Y/N-QUESTION, and BACKCHANNEL. The motivation for conditioning utterance selection on DA label histories is two-fold: first, as we will see in Subsection 4.2, DA labels and human extractive summaries are highly statistically correlated, and thus DA histories can help prediction in the selection task. Second, it is desirable to account for particular dependencies, such as between a WH-QUESTION label and a following STATEMENT, or between a SUGGEST label and a following ACCEPTANCE, in order to create more coherent summaries. For instance, Zechner [2002a] found that a simple technique consisting of linking together questions and answers in conversational speech summaries – and thus preventing the selection of orphan questions or answers – significantly improved their readability according to various human summary evaluations. Question-answer pairs are admittedly not the only discourse relations that we need to preserve in order to create coherent summaries; hence, my aim will be to build techniques to automatically identify such relations and exploit them in utterance selection.

Hidden Markov models (HMMs) are statistical sequential models that have been highly successful in many machine learning applications, especially speech recognition [Rabiner, 1989]. Their underlying assumption that its state variable is influenced by previous values was also found effective in DA classification tasks[Stolcke *et al.*, 2000; Jurafsky, 2003]. I propose to apply their intuition to the summarization case, and model two loosely coupled Markov processes – one binary for summary labels, and one multinomial for DA labels – in a dynamic Bayesian network [Heckerman, 1995].

In subsection 4.1, I will introduce a probabilistic framework that represents DAs and utterance selection labels as two coupled Markov processes, a formulation that will be used in the remaining of the section. I will also discuss various computational models that are applicable to this framework, such as dynamic Bayesian networks (DBNs). Subsection 4.2 characterizes statistical dependencies between extractive summaries and DA labels, which provides empirical grounds for using DA labels. Subsection 4.3 addresses the problem of dynamically determining in the DBN a set of highly predictive parents for each dynamic variable (in particular, the goal is to explicitly model inside the 'switching parent' DBN some dependencies that are representative of various interpersonal behaviors, e.g. speaker-addressee dependencies). Finally, Subsection 4.4 presents various 'oracle' classification results, comparing different feature sets, and contrasting the static case, i.e. standard (non-sequential) classification with no account for DA histories, and the dynamic case, where DA histories are incorporated.

[9] See [Mani and Maybury, 1999] for an overview of the different summarizer categories.

4.1 Probabilistic models

I introduce here a statistical framework under which the above-mentioned dependencies can be modeled, which will help further discussions. Formally, the type of model proposed here and the one used in my preliminary statistical analyzes incorporates two Markov processes: the first chain represents summarization labels $\mathbf{s} =< s_1, ..., s_T >$, where $s_t \in \mathcal{S}$ and $\mathcal{S} = \{\text{SUM}, \text{NONSUM}\}$; it is a set of binary dynamic variables determining whether each of the T utterances is included in the summary or not. The second chain, $\mathbf{d} =< d_1, ..., d_T >$, where $d_t \in \mathcal{D}$ and $\mathcal{D} = \{\text{STATEMENT}, \text{BACKCHANNEL}, ...\}$ is a set of multinomial dynamic variables representing a sequence of DAs assigned to those utterances.

First-order Markov models, i.e. those that simply model the dependency between a state d_t and its predecessor d_{t-1}, are the most common representation of Markov processes in the field. Such models, however, face some limitation in the case of conversational multi-party speech. Indeed, it seems that such models are unable to capture many interesting properties of the inter-personal dynamics of the meeting. First, it would seem reasonable to model the fact that turn or sub-turn units – e.g. utterance, sentences, or DAs – may be overlapping due to competitive speech, and thus that there is no strict ordering between state segments. In this case, it may be useful to condition a given state d_t on the most recent one that ended before d_t itself started. Also, one may want to model the fact that two given states may or may not be contributions of the same speaker.

Finally, two given states d_t and d_{t+k} ($k \geq 2$) may be in certain cases strongly correlated as a result of interpersonal interaction patterns that cannot be captured by first-order HMMs. An example would be the case where d_t corresponds to the current addressee (who made an OFFER in time frame t), and d_{t+k} corresponds to the current speaker (who is making an ACCEPTANCE in time frame $t + k$). It should be noted that, due to the quite disorganized nature of conversational multi-party speech, such cases, where so-called "adjacency" pairs (here, OFFER-ACCEPTANCE) are interleaved with other speech contributions, are quite frequent.

Thus, it seems desirable to be able to dynamically select a set of relevant and predictive parents for each given dynamic variable d_t (or s_t, since the same considerations also apply to summary variables). The three types of dependencies currently modeled in the DBN structure are:

1. **prev**: between d_i and d_{i-1};
2. **ssprev**: between a label and the most recent one of the same speaker (switching);
3. **addrprev**: between a label and the most recent label of the addressee of the current speaker, i.e. a speaker-addressee relationship (switching). This type of dependency of course cannot be trivially determined; subsection 4.3 will describe experiments for identifying addresses.

The mechanism of "switching parents" in DBN was introduced by Bilmes [2000] as dynamic Bayesian multinets. A similar representation was used in [Galley *et al.*, 2004] to perform DA classification using a restricted set of labels {AGREE, DISAGREE, BACKCHANNEL, OTHER}. Its conditional structure was parametrize as a log-linear model and trained with GIS [Darroch and Ratcliff, 1972], and it was shown to outperform a (non-sequential) competitive baseline once switching parent dependencies were added. Regarding the problem of summary label classification, I will postpone the discussion of computational considerations until Section 4.5.

d_i	s_i = SUM	s_i = NONSUM	Odds ration (# YES/# NO)
STATEMENT	2947	26539	.11
BACKCHANNEL	16	15756	.001
DISRUPTION	266	10607	.025
FILLER	14	6618	.002
ACCEPTANCE	124	4740	.026
Y/N-QUESTION	202	3398	.059
SUGGEST	380	2055	.19
UNLABELED	3	1768	.002
WH-QUESTION	105	1074	.098
REJECTION	82	1040	.079
UNCERTAIN	47	789	.059
CHECK	80	593	.14
POLITENESS	10	516	.019
RHETORICAL	16	240	.066
CORRECTION	10	153	.019
OR-CLAUSE	1	122	.008
OPEN-OPTION	15	113	.13
OR-QUESTION	13	112	.12

Table 4: Contingency table of DA and summarization tags.

4.2 Dialog acts and utterance selections: an empirical study

I will try to demonstrate in this subsection that discourse (d_i) and summary labels (s_i) are statistically correlated. To make a meaningful comparison between summarization and DA labels, I mapped the MRDA tagset [Shriberg *et al.*, 2004] from the 1260 unique DA labels actually used in the entire corpus to a set of 18 general labels, such as STATEMENT, QUESTION, BACKCHANNEL, and FILLER.[10] Note that the work presented in [Ji and Bilmes, 2005] reports an accuracy of 66% in DA tagging of the ICSI corpus with a tagset reduced to 62, so it is not unreasonable to assume that we may automatically acquire some of the evidence provided by an even more reduced tagset.

A bivariate analysis revealed, for example, that the summarization and DA dynamic variables s_i and d_i are locally highly correlated. The assumption that the contingency table in Table 4 is uncorrelated is rejected by a χ^2 test ($\chi^2 = 3430, df = 17, p < .001$). Utterances most likely to be included in summaries are statements, suggestions, checks, or-questions, and open-option questions. Those that are the least likely to be included are backchannels, fillers, uncertain utterances, and acceptances. Those kinds of utterances are particularly short. To determine whether length is the primary influence, the same analysis was repeated for utterances longer than 5 seconds, and the correlation is still highly significant ($\chi^2 = 62.39, df = 17, p < .001$). This ruled out almost all backchannels and fillers, but other types of dispreferred utterances such as disruptions and rejections still remained. By further restricting the analysis to a range (length between five and eight seconds), I still found high correlation ($\chi^2 = 49.19, p < .001$), suggesting that DA labels may still be useful predictors even if length is used as a feature.

Other statistical analyzes revealed dynamic dependences between DA and summarization la-

[10]The MRDA tagset, which was used to label the 75 meetings of the ICSI corpus, is a slightly modified version of the DAMSL tagset. It has been adapted to accommodate behaviors particular to meetings. The number of unique tags is particularly high because of the layered labeling scheme of both DAMSL and MRDA that involves assigning one main tag and zero or more secondary tags to each DA unit.

bels; for example, correlation between the utterance of a given speaker and the most recent utterance of her addressee (**addrprev** links) is again highly significant ($\chi^2 = 263.8, \text{df} = 17, p < .001$). The analysis of the contingency table reveals, for example, that utterances elicited by wh- and open-option questions are particularly likely to be included in the summary.

4.3 Inference of dynamic Bayesian networks

The role of the probabilistic model of interpersonal interaction described below is to determine at runtime the structure of the Bayesian network. It links two given dynamic variables at two given points in time (s_i and s_j) if it finds, e.g., that the speaker during s_j addresses the interactant who spoke during s_i. While this model currently only uses speaker-addressee relations to shape the DBN, it is conceptually quite simple to extend it to deal with other types of relations (possibly genre or domain specific), e.g. employer-employee and advisor-student.

4.3.1 Addressee identification

Adjacency pairs (AP) are considered fundamental units of conversational organization [Schegloff and Sacks, 1973]. Their identification is central to our problem, since we need to know the identity of addressees in agreements and disagreements, and adjacency pairs provide a means of acquiring this knowledge. An adjacency pair is said to consist of two parts (later referred to as A and B) that are ordered, adjacent, and produced by different speakers. The first part makes the second one immediately relevant, as a question does with an answer, or an offer does with an acceptance. Extensive work in conversational analysis uses a less restrictive definition of adjacency pair that does not impose any actual adjacency requirement; this requirement is problematic in many respects [Levinson, 1983]. Even when APs are not directly adjacent, the same constraints between pairs and mechanisms for selecting the next speaker remain in place (e.g. the case of embedded question and answer pairs). This relaxation on a strict adjacency requirement is particularly important in interactions of multiple speakers since other speakers have more opportunities to insert utterances between the two elements of the AP construction (e.g. interrupted, abandoned or ignored utterances; backchannels; APs with multiple second elements, e.g. a question followed by answers of multiple speakers).[11]

Information provided by adjacency pairs can be used to identify the target of an agreeing or disagreeing utterance. I define the problem of AP identification as follows: given the second element (B) of an adjacency pair, determine who is the speaker of the first element (A). A quite effective baseline algorithm is to select as speaker of utterance A the most recent speaker before the occurrence of utterance B. This strategy selects the right speaker in 79.8% of the cases in the 50 meetings that were annotated with adjacency pairs. The next subsection describes the machine learning framework used to significantly outperform this already quite effective baseline algorithm.

I view the problem as an instance of statistical ranking, i.e. the problem of selecting, given a set of N possible candidates $\{s_1, ..., s_N\}$ (in our case, potential A speakers), the one candidate s_i that maximizes a given conditional probability distribution.

I use maximum entropy modeling [Berger *et al.*, 1996] to directly model the conditional probability $p(s_i|\mathbf{d})$, where each d_i in $\mathbf{d} = (d_1, ..., d_N)$ is an observation associated with the corre-

[11] The percentage of APs labeled in our data that have non-contiguous parts is about 21%.

sponding speaker s_i. d_i is represented here by only one variable for notational ease, but it possibly represents several lexical, durational, structural, and acoustic observations. Given J feature functions $f_j(\mathbf{d}, s_i)$ and J model parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J)$, the probability of the maximum entropy model is defined as:

$$p_{\boldsymbol{\lambda}}(s_i|\mathbf{d}) = \frac{1}{Z(\mathbf{d})} \exp\left(\sum_{j=1}^{J} \lambda_j f_j(\mathbf{d}, s_i)\right)$$

The only role of the denominator $Z(\mathbf{d})$ is to ensure that $p_{\boldsymbol{\lambda}}$ is a proper probability distribution. To find the most probable speaker of part A, I use the following decision rule:

$$\hat{s} = \underset{s_i \in \{s_1,...,s_N\}}{\arg\max} \left\{ p_{\boldsymbol{\lambda}}(s_i|\mathbf{d}) \right\} = \underset{s_i \in \{s_1,...,s_N\}}{\arg\max} \left\{ \exp\left(\sum_{j=1}^{J} \lambda_j f_j(\mathbf{d}, s_i)\right) \right\}$$

4.3.2 Features for addressee identification

I will now describe the features used to train the maximum entropy model mentioned previously. To rank all speakers (aside from the B speaker) and to determine how likely each one is to be the A speaker of the adjacency pair involving speaker B, I use four categories of features: structural, durational, lexical, and dialog act (DA) information. For the remainder of this section, I will interchangeably use A to designate either the *potential* A speaker or the most recent utterance[12] of that speaker, assuming the distinction is generally unambiguous. I use B to designate either the B speaker or the current spurt for which I need to identify a corresponding A part.

The feature sets are listed in Table 5. Structural features encode some helpful information regarding ordering and overlap of spurts. Note that with only the first feature listed in the table, the maximum entropy ranker matches exactly the performance of the baseline algorithm (79.8% accuracy). Regarding lexical features, I used a count-based feature selection algorithm to remove many first-word and last-word features that occur infrequently and that are typically uninformative for the task at hand. Remaining features essentially contained function words, in particular sentence-initial indicators of questions ("where", "when", and so on).

Note that all features in Table 5 are "backward-looking", in the sense that they result from an analysis of context preceding B. For many of them, I built equivalent "forward-looking" features that pertain to the closest utterance of the potential speaker A that follows part B. The motivation for extracting these features is that speaker A is generally expected to react if he or she is addressed, and thus, to take the fbor soon after B is produced.

4.3.3 Addressee identification: results

I used the labeled adjacency pairs of 50 meetings and selected 80% of the pairs for training. To train the maximum entropy ranking model, I used the generalized iterative scaling algorithm [Darroch and Ratcliff, 1972] as implemented in YASMET.[13]

Table 6 summarizes the accuracy of our statistical ranker on the test data with different feature sets: the performance is 89.39% when using all feature sets, and reaches 90.2% after applying

[12] I build features for both the entire speaker turn of A and the most recent spurt of A.

[13] `http://www.isi.edu/~och/YASMET.html`

Structural features:

· number of speakers taking the floor between A and B
· number of spurts between A and B
· number of spurts of speaker B between A and B
· do A and B overlap?

Durational features:

· duration of A
· if A and B do not overlap: time separating A and B
· if they do overlap: duration of overlap
· seconds of overlap with any other speaker
· speech rate in A

Lexical features:

· number of words in A
· number of content words in A
· ratio of words of A (respectively B) that are also in B (respectively A)
· ratio of content words of A (respectively B) that are also in B (respectively A)
· number of n-grams present both in A and B (I built 3 features for n ranging from 2 to 4)
· first and last word of A
· number of instances at any position of A of each cue word listed in [Hirschberg and Litman, 1994]
· does A contain the first/last name of speaker B?

Table 5: Speaker ranking features.

Feature sets	Accuracy
Baseline	79.80%
Structural	83.97%
Durational	84.71%
Lexical	75.43%
Structural and durational	87.88%
All	89.38%
All (only backward looking)	86.99%
All (Gaussian smoothing, FS)	90.20%

Table 6: Speaker ranking accuracies for different feature sets.

Gaussian smoothing and using incremental feature selection as described in [Berger *et al.*, 1996] and implemented in the yasmetFS package.[14] Note that restricting ourselves to only backward looking features decreases the performance significantly, as we can see in Table 6.

It is overall quite clear that the performance of addressee identification to shape the structure of the DBN is quite good. I now turn to an overall evaluation of the utterance selection system.

4.4 Experiments

Discourse features presented earlier in this section were all tested in a classification task to determine to what extent they are good predictors in the meeting summarization task. I used the

[14] `http://www.isi.edu/~ravichan/YASMET.html`

Durational and acoustic features:

· duration of the utterance
· seconds of silence preceding/during/succeeding the turn
· seconds of silence within the turn (`total_frames - voiced_frames`)
· speech rate
· min/max/mean/median/stddev F0
· pitch range
· min/max/mean/stddev energy
· F0 mean absolute slope

Lexical features:

· number of words in the utterance
· number of content words in the utterance
· number of digits (to identify the so-called digit sections)
· unigram/bigrams/trigrams of the training corpus the most correlated with summarization tags in the development corpus (according to a χ^2 test)

Word statistics:

· max/sum/mean frequency of all terms with in the turn (tf)
· max/sum/mean idf score
· max/sum/mean $tf \cdot idf$ score
· cosine similarity between word vector of utterance with centroid of all word vectors in the meeting/topical segment

Positional and structural features:

· relative position within meeting (1 for first utterance, 0 for last)
· relative position within topical segment
· relative position within speaker turn
· binary variables indicating whether the previous and next turns are of the same speaker
· duration of current topical segment

Discourse features:

· DA label of the utterance
· contextual DA labels; both backward and forward looking features with a 2nd-order Markov assumption on each of the 3 modeled dynamic dependencies (**prev, ssprev, addrprev**). Binary features such as: *the previous utterance of the addressee is a wh-question, the next utterance is a backchannel*.
· same as above, but with a 2nd-order Markov assumption.
· number of seconds between current utterance and all contextual utterances linked to it in the DBN
· number of activation links of each type between current utterance and any number of succeeding ones
· general cues words, as listed in [Hirschberg and Litman, 1994]
· number of pronouns
· number of fillers and fluency devices ("uh", "um", "I mean", etc)
· number of backchannel and acknowledgment tokens ("uh-huh", "ok", "right", etc)
· pronouns, cue words, fillers, backchannel and acknowledgment tokens (each individual token treated as a feature)
· given-new distinction

Table 7: Features for utterance selection.

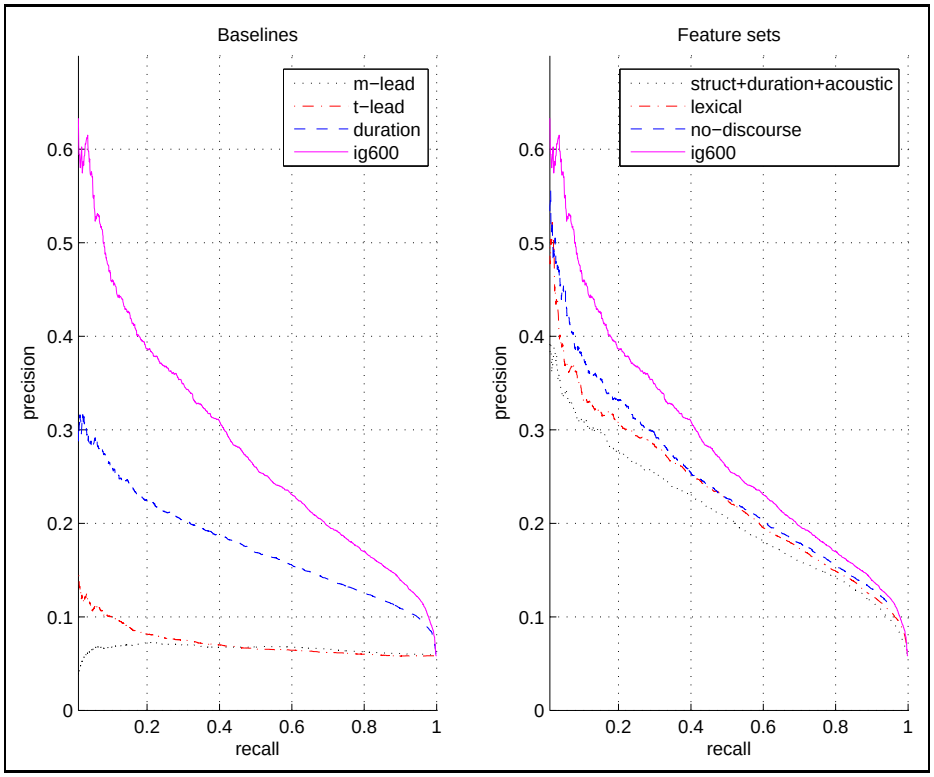

Figure 3: Comparison between the best performing set of features (ig600) and: (a) two baselines: lead of the meeting (m-lead) and lead of each topical segment (t-lead); the best predictor alone (duration). (b) a set of features not depending on lexical information (struct+duration+acoustic), a set of features with only lexical and word frequencies (lexical), the best set of features without discourse features (no-discourse).

Feature set	97.5%	94.2%	90%
baseline: lead of the meeting (m-lead)	.058	.068	.069
baseline: lead of the topic segment (t-lead)	.11	.101	.085
duration	.247	.22	.191
structural+duration+acoustic	.298	.262	.219
lexical+stats	.315	.287	.234
discourse	.251	.227	.199
structural+duration+acoustic+lexical+stats	.339	.297	.235
top 600 information gain features (ig600)	.391	.331	.255
human performance	N/A	.364	N/A

Table 8: Percentages of true positives for different feature sets at given compression lengths. The second column (94.2%) corresponds to the precision/recall break-even point (incidentally, it is an F score).

ground-truth feature values of DAs for the three different predictors discussed in the previous subsection, i.e. **prev**, **ssprev**, and **addreprev**. This is just to assess to potential benefit of DA features, and may of course not be indicative of real performance. I selected 54 meetings that have been annotated with DA tags and used to build extractive summaries. The dataset represents 85,756 instances, of which 4,999 are positive instances (i.e. 5.83%); 4,000 were used for development, 10,000 for testing, and the rest for training. In order to have a basis for comparison, I augmented the discourse-level features discussed previously with features that have been found useful in text summarization tasks: frequency, $tf \cdot idf$, positional features, and so on (see Table 7). I trained models for different groupings of features: structural, lexical, durational, acoustic, global statistics, positional, and discourse. Since features sets were over-generated and contained too many unpredictive and mutually correlated features, I ran on each grouping a correlation-based feature subset selection algorithm based on information gain [Hall, 1998] that seeks to select features that are highly correlated with the class, yet seldom correlated with other features. I used this method to restrict each of the groupings mentioned above to a maximum of 600 features (this had only an impact on sets that contained lexical and discourse features, which are quite numerous). It was interesting to notice that, in the feature set that combined all features (ig600), 84 of the features that were selected are discourse features.

I trained a log-linear model for each feature grouping. Since the data is particularly skewed toward negative examples, no data set outperformed the majority-class baseline in classification accuracy. Performance is summarized in Figure 3 and Table 8 (ig600 refers to the 600 most predictive features across all sets). The main finding provided by table lies in the fact that the ig600 feature set (containing discourse features as well) performs significantly better than the set of all features except discourse (which also contained the same number of features).

Overall, summarization results are admittedly currently quite poor, especially if we consider the fact that best achieved F-measure (.251) used labels that are difficult to obtain automatically. It is however generally agreed that precision/recall evaluation is a too strict for summarization purposes, and doesn't account for cases where two or more utterances may be good alternatives for inclusion in a summary (similar propositional content), and only one of them was included in the gold standard summary. I plan to experiment with automatic evaluation metrics specific to summarization, such as the Pyramid method [Nenkova and Passonneau, 2004] or ROUGE [Lin, 2004].

4.5 Discussion

Regarding the overall structure of the DBN, the goal is of course to perform classification of s_t variables, the DA labels are just additional latent variables that do not directly affect classification evaluation. A main area of future work will be to find efficient structure to predict summary labels, in particular intra-slice structure of the DBN (what are the dependencies between dynamic variables s_t, d_t, and o_t, the observation variable), and its inter-slice structure (i.e., dependencies between dynamic variables at different times). The fact the model contains two Markov processes instead of one does not cause particular difficulties during training (if we still assume that all variables are observables at training time), though training becomes more complex if the DA Markov process is never observed, in which case one solution is to use the EM algorithm [Dempster *et al.*, 1977]. This is a solution that I am planning to investigate if preliminary results with fully observable data (i.e. using human-labeled DAs) appear to be promising. The two advantages of

the EM approach would be: 1) it will not depend on the availability of human-tagged DA labels, which are not necessarily available in corpora other than ICSI. 2) Representations other than DA might be more appropriate in order to model the context that is relevant to summarization. I will finally investigate the problem of exact decoding with the DBN I am proposing, for which I will investigate either approximate solution (e.g. beam search) or exact algorithms. Regarding the latter case, related models that include two or more Markov processes, such as factorial HMM (FHMM) [Ghahramani and Jordan, 1995], have fast decoding algorithms in cases where the number of Markov processes is small.

5 Utterance revision

5.1 Motivation

Purely extractive approaches to summarization may work reasonably well with written texts, but they typically fail to produce good summaries when applied to spoken documents, particularly multi-party speech. Because of their conversational, informal, and competitive nature, transcripts tend to be fragmentary, disfluent, and riddled with errors. These phenomena are particularly prevalent in meetings, since their interactants are generally not professional speakers, as opposed to other speech domains such as broadcast news. As the following reference summary sentence may suggest, it is particularly ineffective to re-use full utterances in a purely extractive approach to summarization:[15]

> Well, I ju- I was just thinking, with reference to uh, things that have [pause] that bear on the content or the status relations, would be the things [pause] without being exhaustive, by any means, but just like I said, if there's a k- a certain topic that comes up in the meeting, and [pause] that knowing their relationship will clarify it, or [pause] if there's a certain dynamic that comes up [pause] so, I mean, a person is asked a whole bunch of questions, more than you'd usually think they'd be asked, and it turns out it's because he's being prepared for a job interview or something like that, then it's useful to know that [pause] that relationship.

A quick look through the summary may convince the reader that such an utterance extract is not clear, concise, and fluent enough to be really useful, even though the fragment was selected by a human.

The goal of the work proposed in this section is twofold: 1) address the above issues by generating summaries that are more coherent, readable, and on-target; 2) provide an effective alternative to utterance-level selection to reduce length. Since full semantic interpretation of unrestricted texts is beyond the state-of-the-art, it is necessary to re-use and modify recognized speech in order to develop robust systems that create such summaries. Certain types of speech fragments are good candidates for revision or deletion:

- **Speech repairs:** only keep the correction in a repair, and drop the reperandum (i.e. the abandoned part of the utterance) and any interleaved edit terms. Note that it is not always possible to handle repairs by removal operations, e.g. "I want to leave from Boston - uh - depart from there", however such cases are quite rare.

[15] This excerpt is taken from one of the 75 extractive summaries created at University of Edinburgh [Carletta *et al.*, 2003] (see Section 2).

- **Parentheticals and fluency devices:** phrases such as “you know”, “I mean”, “basically”, and “anyway” are obvious candidates for deletion. Floor grabbers, fbor holders, holds, and other fi lled pauses such as “um” and “uh” – all frequently appearing in conversational speech – are merely used to take and maintain the fbor, and add little to the propositional content of the utterance.
- **Semantically poor phrases:** many ‘prefabricated phrases’ [Kjellmer, 1991; Nattinger and DeCarrico, 1992] in conversational speech express common concepts that add little to the propositional content of utterances, e.g. “as a matter of fact”, “let’s face it”, “it turns out that”, “like I said”, “some kind of”, and “for some reason”. While such phrases have a role at a more pragmatic level – e.g., as hedging mechanisms (“I would think that”, “it seems like”, “all I am really saying is that”, and “I was just thinking that”), the shade of meaning they introduce is probably too subtle for the kind of processing I am targeting. In other cases, it may be particularly diffi cult to determine the role of such phrases, e.g. general extenders [Overstreet, 1999] can be used as generalization devices (“[...] and stuff like that”), as hedges (“[...], or whatever.”), or might simply represent vague language (“[...] and stuff.”).
- **Discourse markers:** discourse markers such as utterance-initial “so”, “but”, “and”, “anyway”, and “well” have an inter-sentential role and need to be adapted to their new context. While such markers play an important role in earlier stages of the summarization pipeline, and were found helpful predictors in topic segmentation and utterance selection, they are probably best removed at this stage; they may introduce misleading cues regarding the relationship with previous summary sentences.
- **Paraphrasing:** we can exploit the fact that conversational speech has low lexical density [Hughes, 1996], and revise phrases such as “a lot of” into “many”. Paraphrasing can also be used to render spoken utterances into wordings that are more appropriate for written texts.
- **Uninformative constituents:** fi nally, it is of course desirable to go beyond the correction of errors within the summary, and perform summarization below the sentence level by removing the least informative constituents. For this task, it is important to ensure we do not introduce major changes of meaning: at the clause level, adjuncts and optional complements of the verb can be deleted as long as they only contain unimportant information and only shade the meaning of the sentence.

5.2 Related work

A large body of work addresses the problem of revision in text summarization. Text-to-text generation techniques have been applied to single [Jing and McKeown, 2000; Knight and Marcu, 2000; Daumé III and Marcu, 2002; Reizler *et al.*, 2003; Turner and Charniak, 2005] and multi-document summarization [Barzilay *et al.*, 1999; Barzilay and McKeown, 2005], headline generation [Witbrock and Mittal, 1999; Banko *et al.*, 2000; Zajic *et al.*, 2002; Dorr *et al.*, 2003] text compaction for small displays [Corston-Oliver, 2001], text simplifi cation [Chandrasekar *et al.*, 1996; Carroll *et al.*, 1998], information retrieval [Corston-Oliver and Dolan, 1999], audio scanning services for the blind [Grefenstette, 1998], and close captioning [Linke-Ellis, 1999; Robert-Ribes *et al.*, 1999].

A variety of techniques were used for abstraction: information extraction combined with traditional natural language generation techniques [Radev and McKeown, 1998], hand-crafted rules derived from a study of document-abstract pairs [Jing and McKeown, 2000; Dorr *et al.*, 2003],

constituent-removal rules automatically learned from document-abstract pairs [Knight and Marcu, 2002; Reizler *et al.*, 2003], and dynamic (word-level) models such as HMMs [Witbrock and Mittal, 1999; Banko *et al.*, 2000; Zajic *et al.*, 2002].

It is interesting to notice that, besides [Witbrock and Mittal, 1999; Banko *et al.*, 2000; Zajic *et al.*, 2002] all abstraction techniques capitalize on syntactic knowledge. Their case is unusual in the sense that their techniques have only been used to produce headlines, a task where grammaticality is arguably less needed.[16]

Previous work in revision-based summarization of speech transcriptions is becoming relatively large: Zechner's [2002b] revision operations were essentially restricted to disfluency removal (different genres were tested, including meetings). Koumpis and Renals [2003] treated revision as a classification problem, and used lexical and acoustic features to determine which words to keep in short voice messages. Hori [2002] combined different models in a statistical framework, in particular n-gram language models and stochastic dependency grammars. A new DARPA project called EARS (Effective, Affordable, Reusable Speech-to-Text) is quite related to the work cited here, though it does not specifically encompass the problem of summarization. EARS targets the development of speech-to-text technology that produces maximally readable transcriptions, and undoubtedly will stimulate a great deal of future work in revision techniques.

5.3 Overview of proposed work

The proposed utterance revision system is aimed at condensing utterances by removing non-fluencies typical to spontaneous speech, as well as semantically empty phrases ("I mean"), low-content and grammatically optional constituents. It incorporates various knowledge sources: syntactic transformation rules automatically acquired from transcription-abstract parallel texts; a syntax-based language model to promote revision hypotheses that are grammatical; a phrase-based deletion model that was trained to identify phrases likely to be deleted ("you know"); a $tf \cdot idf$ model of word importance, to promote the removal of low information content constituents, and prosodic features to exploit any existing acoustic correlates of words found in abstracts, e.g. prosodically stressed words. The different knowledge sources are integrated in a probabilistic framework where the available evidence of all models is combined to select a globally optimal analysis. A dynamic programming algorithm is used to ensure that the different possible analyzes are scored and ranked in an efficient manner.

More specifically, the proposed utterance revision model is based on general formalisms known as synchronous grammars or transformational grammars, which are designed to generate two languages synchronously (previous work is particularly extensive; see, e.g., [Shieber and Schabes, 1990]). A rule of a synchronous grammar may for example correspond to the deletion of an adverb, as it is the case in Figure 4(3) (more rigorous definitions will be introduced in Section 5.4). Such formalisms have generated substantial interest in other areas of NLP, particularly machine translation, but there are relatively few previous applications to summarization. The proposed model is stochastic and fully trainable from any bi-text of sentence-aligned utterances and revisions, assuming that syntax trees (possibly automatically constructed) are available on both sides.

The advantage of modeling the transformation between a transcription sentence and a revised

[16] Even though headlines have their own grammar, the production of ungrammatical sentences is not particularly harmful since generated sentences are particularly short, e.g. [Banko *et al.*, 2000] produces 5-6 words on average.

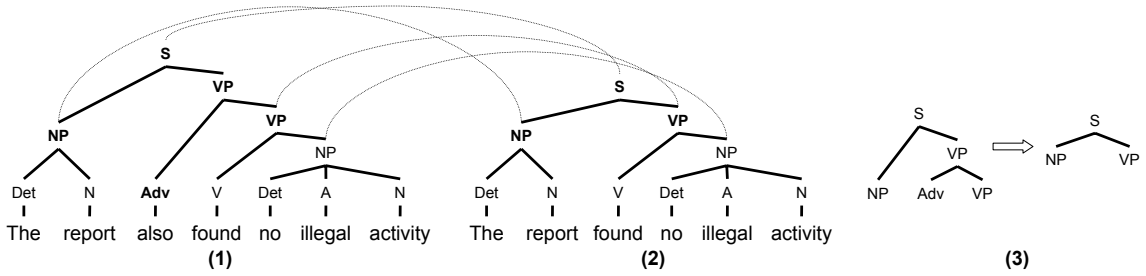

Figure 4: **(1)** and **(2)** show the parse trees of full and compressed sentence. **(3)** an abstraction rule induced by the constituent alignments.

one through a synchronous grammar obviously lies in the grammatical constraints it imposes. These constraints might often be inadequate – since synchronous grammars are automatically induced, and since the application of syntax models to speech is difficult,[17] however it is hard to imagine how purely lexical models (e.g. n-gram models) alone may produce reasonably accurate output in general. For example, while a phrase-based deletion model may be good at finding likely candidates for deletions (e.g. "you know"), it will not particularly work well on an input like "do you know if they left?", and produce a shortened version that may not be appropriately penalized by an n-gram language model.

I anticipate that the proposed work will provide the following research contributions:

- **Better modeling of revision bi-texts (Section 5.4.1):** while the models commonly used in existing systems such as [Jing and McKeown, 2000; Knight and Marcu, 2000; Dorr *et al.*, 2003] take a "parse-and-trim" approach to revision, allowing any subtree to be dropped from the input parse tree, it is clear by looking at existing abstraction training data such as the Ziff-Davis corpus that human abstractors take a quite different approach to revision. Indeed, Knight and Marcu [2000] found, after sentence-aligning documents and abstracts in Ziff-Davis, that it was only possible in 1.7% of the cases to obtain the abstract sentence by simply pruning the parse tree of the corresponding document sentence. To address this modeling problem, I experimented with a representation of revision rules that is more general than the ones in works cited above, since it can model the transformation between any tree pair (while similar approaches have been used in other areas of NLP, e.g. [Shieber and Schabes, 1990], I believe this is new to abstractive summarization).
- **Robust grammar parametrization (Section 5.4.2):** Stochastic synchronous grammars automatically induced from parallel treebanks and used in existing revision systems [Knight and Marcu, 2000; Turner and Charniak, 2005] generally use relative frequencies to assign probabilities to their rules. This represents a major data sparseness issue, since these rules are generally normalized over quite complex low-occurrence structures. While I currently leave aside the problem of parametrizing unrestricted synchronous grammars, I propose a parametrization for the kind of restricted rules (i.e. constituent dropping rules) used in [Knight and Marcu, 2000] and similar work. The parametrization is inspired by [Collins, 1997], and an answer to similar concerns in syntactic parsing. The so-called head-driven

[17] Progress in speech parsing was particularly significant in recent years, e.g. [Charniak and Johnson, 2001] who got 85.3% precision and 86.5% recall, which offers interesting prospects for the application syntax-based revision models in difficult domains such as meetings, even if was applied to accurately transcribed speech.

parametrization introduces independence assumptions, i.e. that the probability of applying a given compression rule is expressed in terms of probabilities of keeping or deleting each syntactic constituent involved in the rule. This dramatically reduces the number of parameters to estimate, and allows the deletion of constituents to be additionally conditioned on lexical information (e.g., this permits the distinction between the adverb "not" – obviously a bad candidate for deletion – and the adverb "actually").

- **Discriminative training framework (Section 5.5):** many additional knowledge sources will be combined in a framework where the scaling factors of the different models are discriminatively trained. Some models, such as the $tf{\cdot}idf$ and the phrase-based deletion model, are quite novel to syntax-driven utterance revision.

5.4 Synchronous grammars for text-to-text generation

The representations of revision rules that have been commonly used in automatically trainable syntax-based sentence condensation systems ([Knight and Marcu, 2000; Knight and Marcu, 2002; Daumé III and Marcu, 2002; Reizler *et al.*, 2003; Turner and Charniak, 2005], which I will henceforth call child-deletion models) are restricted instance of synchronous context free grammars (S-CFG), also known as syntax-directed transduction grammars [Aho and Ullman, 1969] (essentially two context free grammars working in pair), a formalism that can only account for quite limited types of syntactic divergences between tree pairs (in Figure 4, a child-deletion models cannot transform tree (1) into tree (2)). Given their incapacity to model the transformation between a sentence and its revision in most real-world examples, it is arguably desirable to switch to richer models such as synchronous tree substitution grammars (S-TSG) or synchronous tree adjoining grammars (S-TAG), which have been used in other fields of NLP, e.g. [Shieber and Schabes, 1990] in machine translation. These models are generalization of S-CFGs in that rules represent transformations between two arbitrary synchronized trees (Figure 4(3) presents an example of such a rule). The extended domain of locality pertaining to S-TSG allows us to model the transformation between *any* tree pair as a sequence of revision rules, which is computationally appealing since it allows us to exploit any parallel corpus in its entirety (as opposed to only 1.7% of the data corpus such as Ziff-Davis).

I restricted my first experiments to the extraction of S-TSG rules, though I do not exclude the possibility of extracting and using S-TAG rules in the future. Training a synchronous grammar comprises here two subproblems: extract rules from a parallel treebank and assign probabilities to them. These two problems are addressed in the two next subsections (5.4.1 and 5.4.2).

5.4.1 Revision rule extraction

In order to learn revision rules from data, I need corpora of summary sentences paired with the document sentences from which they are derived. Since there is little such data in speech,[18] I will need to acquire or create such data, by asking human annotators to revise (correct speech errors and compress) utterances of the ICSI corpus. So far, I have manually created a small corpus of about 200 sentence pairs, but will need of course to obtain much more data. It might be beneficial to also exploit parallel data of written texts, since they are available in much larger quantities. Even

[18]The abstractive summaries created at the University of Edinburgh might be paired with transcriptions, but the level of abstraction is particularly high, and make this little amount of data particularly difficult to use.

though the two genres are quite different, some abstractive behaviors may be common, and adding textual data may reduce the need for conversational speech data.

Since the amount of available parallel speech data was too small, all preliminary experiments reported here apply to the Ziff-Davis corpus only. In the case of the Ziff-Davis corpus (which is document-aligned but not sentence aligned), I needed first to pair each summary sentence with the document sentence from which it is drawn. There is some significant previous work in sentence alignment for document-abstract pairs [Jing and McKeown, 1999; Daumé III and Marcu, 2004], however I took the relatively simple approach of aligning each abstract sentence to the document sentence that minimizes the word edit distance between the two sentences (I will try more sophisticated sentence alignment methods).

To automatically extract abstraction rules, a first step is to define a constituent alignment between the paired abstract and document sentences, which determine how the two trees are synchronized. While other criteria might be used, I synchronized tree pairs by minimizing the tree edit distance between the two trees [Tai, 1979], i.e. minimize the number of edit operations: constituent and word insertions, substitutions and deletions.[19] The minimization problem was proved to be NP-hard, and I used an approximation algorithm [Zhang and Shasha, 1989] that runs in polynomial time. I set the cost of constituent deletion from full-sentence tree to be zero, since the goal is to identify compressions, and not to penalize them. The tree edit distance provides an appealing generalization of the criterion used to align trees in child-deletion models, since all data exploitable by their method corresponds exactly to the set of tree pairs that have edit distance zero (i.e. any number of constituent deletions, but no insertion or substitution). Allowing the threshold to be set above zero, we can exploit additional data and model transformations that are not purely compressive.[20] I aligned the entire Ziff-Davis corpus using the procedure describe in this section, and found that cost-0 tree pairs account for 1.5% of the data, i.e. 1551 sentence pairs (the Ziff-Davis corpus now represents a larger data set, which explains why the reported percentage is not exactly the same as in [Knight and Marcu, 2000]), cost-1 tree pairs for 3.02%, and cost-2 tree pairs for 4.8%, which in total represents close to 10,000 sentences pairs with relatively low-cost alignments.

Once a constituent alignment is available, it is then trivial to read revision rules off the parallel treebank. From each alignment point such as `S-S` in Figure 4, we search both trees for the next frontier that is only made of other alignment points. In the example, searching from `S-S`, we find the frontier `NP-NP`, `Adv-`ϵ, `VP-VP` (`Adv` is aligned to ϵ, i.e. deleted). Table 9 shows the output of the tree edit distance minimization algorithm on a real example of the Switchboard corpus, and revision rules derived from it. While there are some structural differences between the two trees that are problematic for child deletion models, we can still exploit similarities between both trees to better estimate their probabilities, e.g. all rules in Table except (c) can be used to train such models 9.

[19] Unsynchronized nodes and synchronizations between mismatched node pairs all count as editions. In Figure 4, the only editions are the deletions of a VP, an Adv, and "also" (synchronizations between pre-terminal and leaf nodes are not shown on the figure, to improve readability). .

[20] High thresholds should understandably be avoided, since they allow the alignment of trees that result from bad sentence alignments, or cases where human abstractors took the freedom to significantly reformulate the content of the document. Experimentally, I found that, while the algorithm gives reasonably good alignments with threshold lower than 10, setting the threshold above generally either leads to bad tree alignments or creates sparse alignments, which as we will see later, creates large revision rules that are only sparsely applicable.

f: So I, I don't know in, in terms of other things, other benefi ts other than sort of monetary **c**: I don't know in terms of benefi ts other than monetary		
S$_{1}$**(RB**$_{\epsilon}$ **NP**$_{\epsilon}$ **,**$_{\epsilon}$ **NP**$_{2}$ **VP**$_{3}$**)**	**S**$_{1}$**(NP**$_{2}$ **VP**$_{3}$**)**	(a)
NP$_{2}$(PRP$_{4}$)	NP$_{2}$(PRP$_{4}$)	
PRP$_{4}$(*I*)	PRP$_{4}$(*I*)	
VP$_{3}$(VBP$_{5}$ RB$_{6}$ VP$_{7}$)	VP$_{3}$(VBP$_{5}$ RB$_{6}$ VP$_{7}$)	
VBP$_{5}$(*do*)	VBP$_{5}$(*do*)	
RB$_{6}$(*n't*)	RB$_{6}$(*n't*)	
VP$_{7}$**(VB**$_{8}$ **PP**$_{\epsilon}$ **,**$_{\epsilon}$ **PP**$_{9}$**)**	**VP**$_{7}$**(VB**$_{8}$ **PP**$_{9}$**)**	(b)
VB$_{8}$(*know*)	VB$_{8}$(*know*)	
PP$_{9}$(IN$_{10}$ NP$_{11}$)	PP$_{9}$(IN$_{10}$ NP$_{11}$)	
IN$_{10}$(*in*)	IN$_{10}$(*in*)	
NP$_{11}$(NP$_{12}$ PP$_{13}$)	NP$_{11}$(NP$_{12}$ PP$_{13}$)	
NP$_{12}$(NNS$_{14}$)	NP$_{12}$(NNS$_{14}$)	
NNS$_{14}$(*terms*)	NNS$_{14}$(*terms*)	
PP$_{13}$(IN$_{15}$ NP$_{16}$)	PP$_{13}$(IN$_{15}$ NP$_{16}$)	
IN$_{15}$(*of*)	IN$_{15}$(*of*)	
NP$_{16}$**(NP**$_{\epsilon}$ **,**$_{\epsilon}$ **NP(NP**$_{17}$ **ADVP**$_{18}$**))**	**NP**$_{16}$**(NP**$_{17}$ **ADVP**$_{18}$**)**	(c)
NP$_{17}$**(JJ**$_{\epsilon}$ **NNS**$_{19}$**)**	**NP**$_{17}$**(NNS**$_{19}$**)**	(d)
NNS$_{19}$(*benefits*)	NNS$_{19}$(*benefits*)	
ADVP$_{18}$(ADVP$_{20}$ PP$_{21}$)	ADVP$_{18}$(ADVP$_{20}$ PP$_{21}$)	
ADVP$_{20}$(JJ$_{22}$)	ADVP$_{20}$(JJ$_{22}$)	
JJ$_{22}$(*other*)	JJ$_{22}$(*other*)	
PP$_{21}$(IN$_{23}$ ADJP$_{24}$)	PP$_{21}$(IN$_{23}$ ADJP$_{24}$)	
IN$_{23}$(*than*)	IN$_{23}$(*than*)	
ADJP$_{24}$**(ADVP**$_{\epsilon}$ **JJ**$_{25}$**)**	**ADJP**$_{24}$**(JJ**$_{25}$**)**	(e)
JJ$_{25}$(*monetary*)	JJ$_{25}$(*monetary*)	

Table 9: A tree-to-tree alignment is used to define a derivation involving four revisions (bold-faced rules).

5.4.2 A probabilistic formulation of the sentence revision problem

I discuss here the problem of assigning probabilities to rules discussed in the the previous subsection, and at the end of this section, I will present a lexicalized head-driven parametrization of child-deletion models. I present a probabilistic formulation of utterance condensation and combination where the models are trained from document-abstract pairs, a formulation that draws from previous work based on parallel texts [Witbrock and Mittal, 1999; Banko *et al.*, 2000; Berger and Mittal, 2000; Jing and Hauptmann, 2001; Knight and Marcu, 2002; Reizler *et al.*, 2003]. The overall goal of an utterance revision system is to transform a given full utterance $\mathbf{f} = (f_1, ..., f_n)$ produced by the speech recognizer into a fluent and concise sentence $\mathbf{c} = (c_1, ..., c_m)$. I will first target the problem of sentence or utterance compression, i.e. assume that $\mathbf{c}$ is a subsequence of $\mathbf{f}$. In a noisy-channel framework, which is common to speech and natural language processing tasks ([Jelinek, 1997], *inter alia*), one can use Bayes' theorem to define $\hat{\mathbf{c}}$ as the sentence in the set $\mathbf{C}$ of

all compressions that maximizes the following decision rule:[21]

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c} \in \mathbf{C}} \left\{ Pr(\mathbf{c}|\mathbf{f}) \right\} = \arg\max_{\mathbf{c} \in \mathbf{C}} \left\{ Pr(\mathbf{c}) \cdot Pr(\mathbf{f}|\mathbf{c}) \right\} \quad (2)$$

Equation 2 encompasses our two main problems: the modeling part, i.e. finding linguistically-motivated estimates for $p(\mathbf{c}|\mathbf{f})$ that appropriately capture what lexical material needs to be deleted or revised; finding the string $\hat{\mathbf{c}}$ that maximizes these probability estimates represents our search (or decoding) problem. The modeling problem is further divided into two: language modeling, to ensure that $\hat{\mathbf{c}}$ is well-formed English, and a "transfer model" $p(\mathbf{f}|\mathbf{c})$ to ensure that both $\mathbf{c}$ and $\mathbf{f}$ confer the same meaning.

In child deletion models, the probability estimate of the given transformation between sentences $\mathbf{c}$ and $\mathbf{f}$ (transfer model) is fully determined by their respective parse trees $\pi_{\mathbf{c}}$ and $\pi_{\mathbf{f}}$. If $\tau(\mathbf{f})$ is the set of all parse trees that yield $\mathbf{f}$, we can compute the probability $Pr(\mathbf{f}|\mathbf{c})$ by marginalizing out $\pi_{\mathbf{f}}$ and $\pi_{\mathbf{c}}$:

$$Pr(\mathbf{f}|\mathbf{c}) = \sum_{\pi_{\mathbf{c}} \in \tau(\mathbf{c})} \sum_{\pi_{\mathbf{f}} \in \tau(\mathbf{f})} Pr(\pi_{\mathbf{c}}, \pi_{\mathbf{f}}, \mathbf{f}|\mathbf{c}) = \sum_{\pi_{\mathbf{c}} \in \tau(\mathbf{c})} \sum_{\pi_{\mathbf{f}} \in \tau(\mathbf{f})} Pr(\pi_{\mathbf{c}}, \pi_{\mathbf{f}}|\mathbf{c}) \quad (3)$$

Here, the problem of estimating the probability of $\pi_{\mathbf{f}}$ given $\mathbf{c}$ is generally divided into two: the probability of $\pi_{\mathbf{c}}$ given $\mathbf{c}$ (syntax-based language model) and the probability of the full-sentence parse tree $\pi_{\mathbf{f}}$ given $\pi_{\mathbf{c}} \in \tau(\mathbf{c})$ (the probability assigned by the synchronous grammar):

$$Pr(\mathbf{f}|\mathbf{c}) = \sum_{\pi_{\mathbf{c}} \in \tau(\mathbf{c})} Pr(\pi_{\mathbf{c}}|\mathbf{c}) \cdot \sum_{\pi_{\mathbf{f}} \in \tau(\mathbf{f})} Pr(\pi_{\mathbf{f}}|\pi_{\mathbf{c}}, \mathbf{c}) \quad (4)$$

$$= \sum_{\pi_{\mathbf{c}} \in \tau(\mathbf{c})} Pr(\pi_{\mathbf{c}}|\mathbf{c}) \cdot \sum_{\pi_{\mathbf{f}} \in \tau(\mathbf{f})} Pr(\pi_{\mathbf{f}}|\pi_{\mathbf{c}}) \quad (5)$$

We obtain (4) from (3) using Bayes' rule, and (5) from (4) using $Pr(\mathbf{c}|\pi_{\mathbf{c}}) = 1$. The final expression to maximize is:

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c} \in \mathbf{C}} \left\{ Pr(\mathbf{c}) \cdot \sum_{\pi_{\mathbf{c}} \in \tau(\mathbf{c})} Pr(\pi_{\mathbf{c}}|\mathbf{c}) \cdot \sum_{\pi_{\mathbf{f}} \in \tau(\mathbf{f})} Pr(\pi_{\mathbf{f}}|\pi_{\mathbf{c}}) \right\} \quad (6)$$

Note that, in practice, Equation 6 in [Knight and Marcu, 2000] was approximated by removing the summations over all possible tree pairs, and by using the best hypothesis of a general-purpose syntactic parser as $\hat{\pi}_{\mathbf{f}}$. Their search tries to identify the most likely compression parse tree:

$$\hat{\pi}_{\mathbf{c}} = \arg\max_{\pi_{\mathbf{c}} \in \tau(\mathbf{c})} \left\{ Pr(\mathbf{c}) \cdot Pr(\pi_{\mathbf{c}}|\mathbf{c}) \cdot Pr(\pi_{\mathbf{f}}|\pi_{\mathbf{c}}) \right\} \quad (7)$$

Similarly to PCFG parsing, independence assumptions allow us to factor the probability of the entire structure into probabilities easier to estimate. If Θ is the set of all valid left-most derivations transforming $\pi_{\mathbf{f}}$ into $\pi_{\mathbf{c}}$ through a series of tree transformation operations (i.e. S-TSG rules) such as $\theta_i = t_i^1 \circ t_i^2 \circ ...$, we get:[22]

$$Pr(\pi_{\mathbf{f}}|\pi_{\mathbf{c}}) = \sum_{\theta_i \in \Theta} \prod_{t_i^j \in \theta_i} p(rhs(t_i^j)|lhs(t_i^j)) \quad (8)$$

[21] I will denote general (unknown) probability distributions with $Pr(\cdot)$ and use $p(\cdot)$ for probabilities assigned by the various models.

[22] While S-CFG only allows up to one valid derivation, there can be more than one valid derivation with S-TSG.

5.4.3 Lexicalized, head-driven Parametrization of child-deletion models

Knight and Marcu [2000], among others, relied on a relative frequency estimator for assigning probabilities to S-CFG productions transforming LHS $\rightarrow$ RHS into LHS $\rightarrow$ RHS$'$ (where RHS$'$ was assumed to be a subsequence of RHS). In addition to other problems discussed later, a main issue of their compression model lies in its particularly high number of parameters, which is a problem in their case since the amount of data usable by S-CFG models is quite limited. Similarly to researchers who faced the same problems in other domains (e.g. parsing [Collins, 1997; Charniak, 2000]), I address the sparseness issue by making some independence assumptions regarding context-free (transformation) rules, and assume that the generation of RHS constituents are independent steps. This process, which is now commonly called *horizontal markovization* in the syntactic parsing literature [Klein and Manning, 2003] significantly reduces the number of parameters of the model, and further allows us to lexicalize both syntax-based models (language and compression models) involved in our approach. More specifically, the head-driven language model as defined in[Collins, 1997] breaks down the generation of the RHS as follows:

$$p(\text{RHS}|\text{LHS}) = p(H|P,h) \cdot \prod_{i=1...n+1} p(L_i|P,h,H,C_1^{i-1}) \cdot \prod_{i=1...m+1} p(R_i|P,h,H,C_1^{i-1})$$

where P is the parent (i.e. LHS), H is the head constituent of the RHS, h the head word, L_i and R_i are the m left and n right modifiers and their head words.[23] One key factor lies in C_i^{i-1}, the left and right contexts between the modifier and the head constituents: they need to capture relevant properties of the parse tree to make the assumption of independence between generated constituents linguistically plausible (for example, model 1 in [Collins, 1997] represents C_i^{i-1} as the distance between H and the modifier, i.e. $C_1^{i-1} = i - 1$).

One of the contributions of our proposed work will lie in the application of a head-driven lexicalized parametrization proposed in [Collins, 1997] to sentence compression. This provides both the benefit of reducing the sparseness problem in the compression model, and of giving more reasonable lexicalized estimates (than [Knight and Marcu, 2000]) for the probability of deleting certain constituents: for example, the verb phrase "are not leaving", which is here the yield of the production VP $\rightarrow$ VBP RB VP, cannot be effectively differentiated from e.g. "are actually leaving" by the compression model used in [Knight and Marcu, 2000], since the parse trees of the two phrases only differ by their leaves. All evidence emerging from parallel summarization corpora indicates that "not" is not a word any abstractor would want to delete; however, the unlexicalized model can only collect statistics regarding syntactic labels such as RB (adverb), and is unable to learn both preferences for not deleting "not" and for deleting "actually".[24]

Similar to the syntax-based language model case, we make some independence assumptions between constituents and break down the probability of transforming RHS' into RHS (given LHS)

[23] L_{n+1} and R_{m+1} are special STOP symbols to indicate where modifiers end on both sides.

[24] This particular problem can't be solved by using a finer-grained set syntactic categories, i.e. one that distinguishes "not" from other adverbs, since one cannot enumerate all cases where lexical distinctions may be helpful. Consider for example the noun phrase "the repossessed real estate", headed by the production NP $\rightarrow$ DT JJ JJ NN. While it is arguably preferable to delete "repossessed" instead of "real", an unlexicalized model is unable to make a distinction between the two.

into a product of probabilities:

$$p(\text{LHS} \rightarrow \text{RHS}'|\text{LHS} \rightarrow \text{RHS}) = \prod_{\substack{i=1\ldots n \\ L_i(l_i)\in\text{RHS}}} p(L_i(l_i)|P,h,H,C_1^{i-1}) \cdot \prod_{\substack{i=1\ldots m \\ R_i(r_i)\in\text{RHS}}} p(R_i(r_i)|P,h,H,C_1^{i-1})$$

Each element L_i and R_i in the above equation appears both in RHS' and RHS and is thus not deleted, and our goal will be to determine what needs to be modeled about context in the parse tree (C_1^{n-1}) so that the constituent deletion probability can reflect what is seen in the data. For example, the deletion of a "not" in the above example is conditioned on the fact that it is a right modifier (R_i), and that $P = \text{VP}$, $h = \text{is}$, $H = \text{VBP}$.

5.5 Additional knowledge sources

So far, I have discussed how to build the transfer model and the language model for the summary sentence. Another component of my proposed work is an extension of our probabilistic framework so that it can incorporate many different knowledge sources, and can be directly trained to best discriminate between positive evidence (reference compressions) and negative evidence (compressions we want our system to avoid producing). This framework, commonly called n-best re-ranking or discriminative training, has been successfully applied in other tasks, such as speech recognition and machine translation [Bahl *et al.*, 1986; Och and Ney, 2002]. More specifically, we can use an n-best list $(\pi_{c_1}, ..., \pi_{c_n})$ of a base compression system as one trained with relative frequencies (Equation 3), and learn to identify within the n-best list the good (reference) compressed tree using statistical ranking. Using the maximum entropy framework [Berger *et al.*, 1996], the decision rule (to find the most probable parse, as in Equation 7) becomes:

$$\hat{\pi}_{\mathbf{c}} = \underset{\pi_{\mathbf{c_i}}\in\{\pi_{\mathbf{c}_1},\ldots,\pi_{\mathbf{c}n}\}}{\arg\max} \left\{ \exp\left(\sum_{j=1}^{J} \lambda_j h_j(\pi_{c_i}, \pi_{\mathbf{f}}) \right) \right\} \tag{9}$$

$\left\{h_j\right\}_{j=1}^{J}$ are features observed from sentence-compression tree pairs $(\pi_{c_i}, \pi_{\mathbf{f}})$ and their yields, and λ_j are their corresponding weights. Since we can use probabilities such as the transfer-model score between $\mathbf{c_i}$ and $\mathbf{f}$ as features, this incidentally means that the above expression is a generalization of the decision criterion used in previous work such as [Knight and Marcu, 2002].[25] This generalization brings an important advantage over the model expressed in Equation 7: it is not limited to using only three information sources, and any predictor that is deemed useful in the compression task can be incorporated and have its weight optimized according to the training data.

Certain features have been shown in the summarization field to be important when building sentence extraction systems for summarization. For example, $tf \cdot idf$ [Salton and Buckley, 1988], and acoustic features [Koumpis and Renals, 2003; Inoue *et al.*, 2004; Maskey and Hirschberg, 2005]. I will experiment with other features, in particular phrase-based deletion models.

[25] Indeed, if we use the three probability models listed in Equation 7 as features, and set $h_1(\pi_{c_i}, \pi_{\mathbf{f}}) = \log p(\mathbf{c}_i)$, $h_2(\pi_{c_i}, \pi_{\mathbf{f}}) = \log p(\pi_t|\pi_{c_i})$, $h_3(\pi_{c_i}, \pi_{\mathbf{f}}) = \log p(\pi_{c_i}|\mathbf{c}_i)$, and all parameters λ_j to 1, then the decision rules in Equations 7 and 9 are equivalent.

5.6 Search

The final problem to address is to find the most likely revision among $\mathbf{C}$ given a transcription sentence $\mathbf{c}$. There are three different search problems: most probable revision $\hat{\mathbf{c}}$ (MPR), most probable parse $\hat{\pi}_{\mathbf{c}}$ (MPP), and most probable derivation $\hat{\delta}$ (MPD). In the simple case node-deletion revision models, the MPP is equivalent to the MPD, and the decoding problem can be solved by augmenting an existing probabilistic lexicalized parser by incorporating synchronous grammar scores discussed in Section 5.4.3. There is an overhead in complexity: if k is the maximum fan-out among all CFG rules, then for each chart item, we have to consider $O(2^k)$ possible configurations of deletions/non-deletions of constituents in the RHS. While this does not depend on the length of the input sentence, it is likely to be computationally too prohibitive. An approximate alternative is to extract a large n-best list (or use directly the pruned derivation forest encoded in its chart) to restrict the search space. The latter methods are also applicable to MPR decoding, since summing together the scores of revision parses that yield the same string within a large n-best list is an approximate way of finding $\hat{\mathbf{c}}$.

I will not discuss here the decoding with more general types of synchronous grammars (S-TSG) [Poutsma,] and efficiently combining synchronous grammars with n-gram models [Huang *et al.*, 2005], though I will investigate these problems in my thesis.

6 Plan for completion

Topic segmentation

I already implemented a baseline topic segmentation system [Galley *et al.*, 2003], using an off-the-shelf learning algorithm [Quinlan, 1993] to train a binary classifier exploiting lexical, durational, and distributional features. I propose the following extensions to the above work:

- Study the acoustic correlate of topic shifts by constructing acoustic and prosodic features for the classifier **[July 2006]**;
- Address the issues discussed in Section 3.6 regarding the training of segmenters; devise techniques to minimize P_k more directly. **[July-Aug 2006]**.

Utterance selection

A baseline utterance selection system that uses lexical, distributional (e.g. $tf \cdot idf$), durational, positional, and acoustic features is already available (Section 4.4). The probabilistic model of interpersonal interaction (speaker-addressee link detection) presented in Section 4.3 is also already implemented. Remaining work will focus on effective modelling of inter-sentential (discourse-level) influences across meetings.

- Implement and train the DBN network described in Section 4.1, possibly making use existing tools such as the Graphical Model ToolKit (GMTK); [Bilmes, 2003] **[Nov 2005]**.
- Model selection: experiment with different network structures, including discriminative and generative structures. Comparison against simpler dynamic models, such as simple HMMs **[Nov 2005]**.

- Discover latent variables that might be more appropriate in the task; perform EM training with fully hidden latent variables, and compare performance against DA models **[Nov-Dec 2005]**.
- Research and evaluate other interpersonal relations (other than speaker-addressee), e.g. models that depends on social status (employer-employee, advisor-student, etc.) **[Jan 2006]**.

Utterance revision

The bulk of my remaining thesis work will lie in the utterance revision component. A baseline system that essentially reproduces the work of [Knight and Marcu, 2000] is already implemented, as well as several other models used during decoding (e.g. $tf \cdot idf$), and the head-driven decomposition of compression rules of Section 5.4.2.

- Data acquisition: select 2000-5000 meeting average-length extract utterances; I (and possibly other human annotators) will produce abstracts for them **[Nov 2005-August 2006]**.
- Model training for S-CFG rules: implement and incorporate all models described in Section 5.5 **[Feb 2006]**.
- Decoder: use a publicly available lexicalized PCFG parser (e.g. [Collins, 1999; Bikel, 2004]) and incorporate the parametrization of the child-deletion model described in Section 5.4.3, in order to extract the most probable revision. Examine the use of alternatives, such as n-best list extraction or direct processing of the parse chart **[Mar-Apr 2006]**.
- Parameter tuning: train the scaling factors of the different models used in revision (revision rules, n-gram language model, $tf \cdot idf$, etc) **[May 2006]**.
- Experiments with speech data: compare in particular output from human transcription and from ASR output. Application of the tree alignment algorithm to the speech data. **[Jun 2006]**.
- If time permits, handle revision rules that are more general than child-deletion rules, and find robust estimates to compute their probabilities. Adapt existing S-TSG parsing algorithms.

Evaluation, system integration, and thesis writing

- Evaluate each three components under standard conditions, e.g. human vs. ASR transcription. **[Sept-Nov 2006]**.
- System integration, and involvement of a large number of human subjects for a qualitative evaluation of summaries produced under various conditions: summaries topically organized or not; baseline or DBN-based utterance selection; utterance revision or extracts **[Sept-Nov 2006]**.
- Thesis writing **[Sept-Jan 2006]**.
- Thesis defense **[Feb 2007]**.

Bibliography

[Aho and Ullman, 1969] Alfred V. Aho and Jeffrey D. Ullman. Properties of syntax directed translations. *Journal of Computer and System Sciences*, 3(3):319–334, 1969.

[Allan *et al.*, 1998] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[Allen and Core, 1997] James Allen and Mark Core. Draft of DAMSL: Dialog Act Markup in Several Layers. Unpublished manuscript, 1997.

[Bahl *et al.*, 1986] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 49–52, Tokyo, April 1986.

[Banko *et al.*, 2000] Michele Banko, Vibhu Mittal, and Michael Witbrock. Headline Generation Based on Statistical Translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL–2000)*, pages 318–325, Hong Kong, China, October 1–8 2000.

[Barzilay and Elhadad, 1997] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proc. of the Intelligent Scalable Text Summarization Workshop (ISTS), ACL*, Madrid, Spain, 1997.

[Barzilay and McKeown, 2005] Regina Barzilay and Kathleen R. McKeown. Sentence Fusion for Multidocument Summarization. *Computational Linguistics (to appear)*, 2005.

[Barzilay *et al.*, 1999] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL–99)*, pages 550–557, University of Maryland, June 20–26 1999.

[Beeferman *et al.*, 1997] Doug Beeferman, Adam Berger, and John Lafferty. Text Segmentation Using Exponential Models. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 35–46. Association for Computational Linguistics, Somerset, New Jersey, 1997.

[Beeferman *et al.*, 1999] D. Beeferman, A. Berger, and J. Lafferty. Statistical Models for Text Segmentation. *Machine Learning*, 34(1–3):177–210, 1999.

[Berger and Mittal, 2000] Adam Berger and Vibhu Mittal. Query-Relevant Summarization Using FAQs. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL–2000)*, pages 294–301, Hong Kong, October 1–8 2000.

[Berger *et al.*, 1996] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, 1996.

[Bikel, 2004] Dan Bikel. *On the Parameter Space of Generative Lexicalized Statistical Parsing Models*. PhD thesis, University of Pennsylvania, 2004.

[Bilmes, 2000] Jeffrey Bilmes. Dynamic Bayesian Multinets. In *Proc. of UAI–2000*, 2000.

[Bilmes, 2003] Jeffrey Bilmes. Graphical Models and Automatic Speech Recognition. In *Mathematical Foundations of Speech and Language Processing*. Springer-Verlag, 2003.

[Boguraev and Kennedy, 1997] Branimir Boguraev and Christopher Kennedy. Salience-based content characterisation of text documents. In *Proc. of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 2–9, Madrid, Spain, 1997.

[Burger *et al.*, 2002] Susanne Burger, Victoria MacLaren, and Hua Yu. The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. In *Proc. of the International Conference on Spoken Language Processing (ICSLP–2002)*, Denver, CO, USA, 2002.

[Carletta *et al.*, 2003] Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holger Voormann. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363, 2003.

[Carletta *et al.*, 2005] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI Meeting Corpus: A Pre-Announcement. In *Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, 2005.

[Carroll *et al.*, 1998] John Carroll, Guidon Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998.

[Chandrasekar *et al.*, 1996] Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. Motivations and Methods for Text Simplification. In *Proc. of the 16th International Conference on Computational Linguistics (COLING–96)*, pages 1041–1044, 1996.

[Charniak and Johnson, 2001] Eugene Charniak and Mark Johnson. Edit Detection and Parsing for Transcribed Speech. In *Proc. of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL–01)*, pages 118–126, 2001.

[Charniak, 2000] Eugene Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL–2000)*, pages 132–139, Seattle, Washington, April 29 – May 3 2000.

[Choi, 2000] Freddy Y.Y. Choi. Advances in domain independent linear text segmentation. In *Proc. of NAACL'00*, 2000.

[Collins, 1997] Michael Collins. Three Generative, Lexicalized Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL–97)*, pages 16–23, Madrid, Spain, July 7-12 1997.

[Collins, 1999] Michael Collins. *Head-driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, Philadelphia, 1999.

[Corston-Oliver and Dolan, 1999] Simon H. Corston-Oliver and William B. Dolan. Less is more: Eliminating Index Terms From Subordinate Clauses. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL–99)*, pages 349–356, University of Maryland, MD, June 20–26 1999.

[Corston-Oliver, 2001] Simon H. Corston-Oliver. Text Compaction for Display on Very Small Screens. In *Proc. of the Workshop on Automatic Summarization at the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 89–98, 2001.

[Darroch and Ratcliff, 1972] J. N. Darroch and D. Ratcliff. Generalized Iterative Scaling for Log-Linear Models. *Annals of Mathematical Statistics*, 43(5):1470–1480, October 1972.

[Daumé III and Marcu, 2002] Hal Daumé III and Daniel Marcu. A Noisy-Channel Model for Document Compression. In *Proc. of the Conference of the Association of Computational Linguistics (ACL–02)*, 2002.

[Daumé III and Marcu, 2004] Hal Daumé III and Daniel Marcu. A Phrase-Based HMM Approach to Document/Abstract Alignment. In *Proceedings of EMNLP*, Barcelona, Spain, 2004.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(Ser B):1–38, 1977.

[Dorr *et al.*, 2003] Bonnie J. Dorr, David Zajic, and Richard Schwartz. Hedge: A Parse-and-Trim Approach to Headline Generation. In *Proc. of the HLT-NAACL Text Summarization Workshop and Document Understanding Conference (DUC–2003)*, pages 1–8, Edmonton, Canada, 2003.

[Galley *et al.*, 2003] Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse Segmentation of Multi-party Conversation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–03)*, pages 562–569, Sapporo, Japan, July 2003.

[Galley *et al.*, 2004] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL–04)*, pages 669–676, Barcelona, Spain, July 2004.

[Ghahramani and Jordan, 1995] Zoubin Ghahramani and Michael I. Jordan. Factorial Hidden Markov Models. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8, pages 472–478. MIT Press, 1995.

[Grefenstette, 1998] Gregory Grefenstette. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Proceedings of the AAAI Spring Workshop on Intelligent Text Summarization*, pages 111–115, Stanford University, CA, March 23-25 1998.

[Grosz and Sidner, 1986] Barbara J. Grosz and Candace L. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, July–September 1986.

[Hahn and Reimer, 1999] Udo Hahn and Ulrich Reimer. Knowledge-Based Text Summarization: Salience and Generalization Operators for Knowledge Base Abstraction. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 215–232. MIT Press, July 1999.

[Hall, 1998] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, 1998.

[Hearst, 1994] Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL–94)*, pages 9–16, Las Cruces, New Mexico, June 27–30 1994.

[Heckerman, 1995] David Heckerman. A Tutorial on Learning With Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, WA, 1995.

[Hirschberg and Litman, 1994] Julia Hirschberg and Diane J. Litman. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530, 1994.

[Hirschberg and Nakatani, 1996] Julia Hirschberg and Christine H. Nakatani. A prosodic analysis of discourse segments in direction-given monologues. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL–96)*, pages 286–293, Santa Cruz, California, June 24–27 1996.

[Hori, 2002] Chiori Hori. *A Study on Statistical Methods for Automatic Speech Summarization*. PhD thesis, Tokyo Institute of Technology, March 2002.

[Huang *et al.*, 2005] Liang Huang, Hao Zhang, and Dan Gildea. Machine Translation as Lexicalized Parsing with Hooks. In *Proceedings of the 9th International Workshop on Parsing Technologies (IWPT–05)*, 2005.

[Hughes, 1996] Rebecca Hughes. *English in Speech and Writing: investigating language and literature*. Routledge publishing, 1996.

[Inoue *et al.*, 2004] Akira Inoue, Takayoshi Mikami, and Yoichi Yamashita. Improvement of Speech Summarization Using Prosodic Information. In *Proc. of Speech Prosody*, 2004.

[Janin *et al.*, 2003] Adam Janin, Dan Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI Meeting Corpus. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP–03)*, Hong Kong, 2003.

[Jelinek, 1997] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1997.

[Ji and Bilmes, 2005] Gang Ji and Jeffrey Bilmes. Dialog act tagging using graphical models. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP–05)*, 2005.

[Jing and Hauptmann, 2001] Rong Jing and Alexander Hauptmann. Title Generation for Machine-Translated Documents. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence (IJCAI–01)*, Seattle, WA, August 4–10 2001.

[Jing and McKeown, 1999] Hongyan Jing and Kathleen R. McKeown. The Decomposition of Human-Written Summary Sentences. In *Proceedings of the 22nd Conference on Research and Development in Information Retrieval (SIGIR–99)*, Berkeley, CA, August 15–19 1999.

[Jing and McKeown, 2000] Hongyan Jing and Kathleen R. McKeown. Cut and Paste Based Text Summarization. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics NAACL-2000*, pages 178–185, Seattle, Washington, April 29 – May 3 2000.

[Jurafsky, 2003] Dan Jurafsky. Pragmatics and Computational Linguistics. In Laurence R. Horn and Gregory Ward, editors, *Handbook of Pragmatics*. Blackwell, Oxford, UK, 2003.

[Kjellmer, 1991] Goran Kjellmer. A mint of phrases. In Karin Aijmer and Bengt Altenberg, editors, *English Corpus Linguistics*, pages 111–127. Longman, 1991.

[Klein and Manning, 2003] Dan Klein and Christopher D. Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–03)*, 2003.

[Knight and Marcu, 2000] Kevin Knight and Daniel Marcu. Statistics-Based Summarization — Step One: Sentence Compression. In *The 17th National Conference on Artificial Intelligence (AAAI–2000)*, pages 703–710, Austin, TX, July 30th – August 3rd 2000.

[Knight and Marcu, 2002] Kevin Knight and Daniel Marcu. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. *Artificial Intelligence*, 139(1), 2002.

[Koumpis and Renals, 2003] K. Koumpis and S. Renals. Evaluation of extractive voicemail summarization. In *Proc. ISCA Workshop on Multilingual Spoken Document Retrieval*, pages 19–24, 2003.

[Kupiec *et al.*, 1995] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington, 1995. ACM Press.

[Levinson, 1983] Stephen C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

[Lin, 1991] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, 37(1):145–151, 1991.

[Lin, 2004] Chin-Yew Lin. ROUGE: a package for automatic evaluation of summaries. In *Proc. of workshop on text summarization, ACL-04*, 2004.

[Linke-Ellis, 1999] Nanci Linke-Ellis. Closed Captioning in America: Looking Beyond Compliance. In *Proceedings of the TAO Workshop on TV Closed Captions for the hearing impaired people*, pages 43–59, Tokyo, Japan, November 20 1999.

[Litman and Passonneau, 1995] Diane J. Litman and Rebecca J. Passonneau. Combining Multiple Knowledge Sources for Discourse Segmentation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, pages 108–115, Cambridge, Massachusetts, June 26-30 1995.

[Mani and Maybury, 1999] Inderjeet Mani and Mark T. Maybury. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA, 1999.

[Marcu, 1997] Daniel Marcu. From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 11 1997.

[Maskey and Hirschberg, 2005] Sameer Maskey and Julia Hirschberg. Comparing Lexial, Acoustic/Prosodic, Discourse and Structural Features for Speech Summarization. In *Proc. of Eurospeech*, 2005.

[McCowan *et al.*, 2003] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling Human Interaction in Meetings. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP–03)*, Hong Kong, April 2003.

[McKeown *et al.*, 2005] Kathleen McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey. From Text Summarization to Speech Summarization. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP–05), Special session on Human Language Technology: Applications and Challenges for Speech Processing.*, 2005.

[Mirghafori *et al.*, 2004] Nikki Mirghafori, Andreas Stolcke, Chuck Wooters, Tuomo Pirinen, Ivan Bulyko, Dave Gelbart, Martin Graciarena, Scott Otterson, Barbara Peskin, and Mari Ostendorf. From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System. In *Proc. of the International Conference of Spoken Language Processing (ICSLP–04)*, Jeju, Korea, 2004.

[Nattinger and DeCarrico, 1992] James Nattinger and Jeanette DeCarrico. *Lexical Phrases and Language Teaching*. Oxford University Press, 1992.

[Nelder and Mead, 1965] J. Nelder and R. Mead. A simplex method for function minimization. In *Computer Journal*, number 7, pages 308–313, 1965.

[Nenkova and Passonneau, 2004] Ani Nenkova and Rebecca Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[NIST, 2004a] NIST. NIST Pilot Meeting Corpus, 2004. `http://www.nist.gov/speech/test_beds/mr_proj/meeting_corpus_1`.

[NIST, 2004b] NIST. NIST Rich Transcription 2004 Evaluation (RT–04), 2004. `http://nist.gov/speech/tests/rt`.

[Och and Ney, 2002] Franz Josef Och and Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, PA, July 2002.

[Overstreet, 1999] Maryann Overstreet. *Whales, Candlelight, and Stuff Like That: General Extenders in English Discourse*. Oxford University Press, 1999.

[Passonneau and Litman, 1993] Rebecca J. Passonneau and Diane J. Litman. Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proc. of the ACL*, 1993.

[Passonneau and Litman, 1997] Rebecca J. Passonneau and Diane J. Litman. Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, 23(1):103–139, 1997.

[Pevzner and Hearst, 2002] L. Pevzner and M. Hearst. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28 (1):19–36, 2002.

[Poutsma,] Arjen Poutsma. Data-Oriented Translation. In *Proc. of COLING–2000*, pages 635–641.

[Powell, 1964] M.J.D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comput. J.*, 7(2):155–162, 1964.

[Quinlan, 1993] Ross Quinlan. *C4.5: Programs for Machine Learning*. Machine Learning. Morgan Kaufmann, 1993.

[Rabiner, 1989] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recogntion. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[Radev and McKeown, 1998] Dragomir Radev and Kathleen R. McKeown. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, 24(3), 1998.

[Reizler *et al.*, 2003] Stefan Reizler, Tracy H. King, Richard Crouch, and Annie Zaenen. Statistical Sentence Condensation using Ambiguity Packing and Stochastic Disambiguation Methods for Lexical-Functional Grammar. In *Proceedings of HLT/NAACL*, pages 197–204, 2003.

[Renals and Ellis, 2003] S. Renals and D. Ellis. Audio information access from meeting rooms. In *Proc. IEEE ICASSP*, volume 4, pages 744–747, 2003.

[Reynar, 1998] Jeffrey Reynar. *Topic Segmentation: Algorithms and Applications*. PhD thesis, University of Pennsylvania, 1998.

[Robert-Ribes *et al.*, 1999] J. Robert-Ribes, S. Pfeiffer, R. Ellison, and D. Burnham. Semi-automatic captioning of TV programs, an Australian perspective. In *Proceedings of the TAO Workshop on TV Closed Captions for the Hearing Impaired People*, pages 87–100, Tokyo, Japan, 1999.

[Salton and Buckley, 1988] Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[Schegloff and Sacks, 1973] Emanuel A. Schegloff and Harvey Sacks. Opening up closings. *Semiotica*, 7-4:289–327, 1973.

[Searle, 1969] John R. Searle. *Speech Acts: an essay in the philosophy of language*. Cambridge University Press, 1969.

[Shieber and Schabes, 1990] Stuart M. Shieber and Yves Schabes. Synchronous Tree-Adjoining Grammars. In *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 253–258, Helsinki, Finland, 1990.

[Shriberg *et al.*, 2004] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *SIGdial Workshop on Discourse and Dialogue*, pages 97–100, 2004.

[Spärck Jones, 1993] Karen Spärck Jones. What might be in a Summary? In Knorz, Krause, and Womser-Hacker, editors, *Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26, Konstanz, DE, 1993. Universitätsverlag Konstanz.

[Stolcke *et al.*, 2000] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373, 2000.

[Tai, 1979] Kuo-Chung Tai. The Tree-to-Tree Correction Problem. *Journal of the ACM*, 26(3):422–433, 1979.

[Tür *et al.*, 2001] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57, 2001.

[Turner and Charniak, 2005] Jenine Turner and Eugene Charniak. Supervised and Unsupervised Learning for Sentence Compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL–2005), to appear*, 2005.

[Utiyama and Isahara, 2001] M. Utiyama and H. Isahara. A Statistical Model for Domain-Independent Text Segmentation. In *Proc. of the ACL*, 2001.

[Waibel *et al.*, 2001] Alex Waibel, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner. Advances in automatic meeting record creation and access. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP–01)*, Salt Lake City, UT, May 2001.

[Witbrock and Mittal, 1999] Michael J. Witbrock and Vibhu O. Mittal. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR–99), Poster Session*, pages 315–316, Berkeley, CA, August 1999.

[Wright, 1996] M.H. Wright. Direct search methods: Once scorned, now respectable. In D.F. Griffiths and G.A. Watson, editors, *Numerical Analysis 1995, Proceedings of the 1995 Dundee Biennial Conference in Numerical Analysis*, pages 191–208. Addison-Wesley Longman, Harlow, UK, 1996.

[Zajic *et al.*, 2002] David Zajic, Bonnie J. Dorr, and Richard Schwartz. Automatic Headline Generation for Newspaper Stories. In *In Proceedings of the ACL Workshop on Text Summarization (DUC–2002)*, pages 78–85, Philadelphia, PA, 2002.

[Zechner and Waibel, 2000] Klaus Zechner and Alex Waibel. DiaSumm: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains. In *Proceedings of COLING-2000, Saarbruecken, Germany*, pages 968–974, 2000.

[Zechner, 2002a] Klaus Zechner. Automatic Summarization of Open Domain Multi-Party Dialogues in Diverse Genres. *Computational Liguistics, Special Issue on Summarization*, 28(4):447–485, 2002.

[Zechner, 2002b] Klaus Zechner. Summarization of Spoken Language - Challenges, Methods, and Prospects. In *Speech Technology Expert magazine*, volume Issue 6, January 2002.

[Zhang and Shasha, 1989] K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18(6):1245–1262, 1989.