SPECOM 1898
17 June 2010 Disk Used
ARTICLE IN PRESS
No. of Pages 15, Model 5+

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2010) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Long story short – Global unsupervised models for keyphrase based meeting summarization

Korbinian Riedhammer [a,b,*], Benoit Favre [c,b], Dilek Hakkani-Tür [b]

[a] Lehrstuhl für Informatik 5, Universität Erlangen-Nürnberg, Martensstraße 3, 91058 Erlangen, Germany
[b] International Computer Science Institute, 1947 Center St, Berkeley, CA 94704, USA
[c] Laboratoire d'Informatique de l'Universite du Maine, Avenue Laënnec, 72085 Le Mans CEDEX 9, France

Received 12 October 2009; received in revised form 2 June 2010; accepted 2 June 2010

Abstract

We analyze and compare two different methods for unsupervised extractive spontaneous speech summarization in the meeting domain. Based on utterance comparison, we introduce an optimal formulation for the widely used greedy maximum marginal relevance (MMR) algorithm. Following the idea that information is spread over the utterances in form of concepts, we describe a system which finds an optimal selection of utterances covering as many unique important concepts as possible. Both optimization problems are formulated as an integer linear program (ILP) and solved using public domain software. We analyze and discuss the performance of both approaches using various evaluation setups on two well studied meeting corpora. We conclude on the benefits and drawbacks of the presented models and give an outlook on future aspects to improve extractive meeting summarization.

Keywords: Multi-party meetings speech; Summarization; Keyphrases; Global optimization

1. Introduction

Wherever people work together, there are (regular) meetings to check on the current status, discuss problems or outline future plans. Recording these get-togethers is a good way of documenting and archiving the progress of a group. This can be done for example by a distant microphone on a table or by integrating a storage device in a teleconference system. Once acquired, these data can serve several purposes: Non-attendants can go through the meeting to get up to date on group discussions, or participants can check certain points of the agenda in case of uncertainty or lack of notes. However, listening to the whole meeting is tedious and one should be able to directly access the relevant information.

Automatic meeting summarization is one step towards the development of efficient user interfaces for accessing meeting archives. In this work, we study the selection of a concise set of relevant utterances[1] in meeting transcripts generated by automatic speech recognition (ASR). The selected meeting extracts can then either be juxtaposed to form a short text summarizing a meeting or used as a starting point to enhance browsing experience.

Extractive summarization algorithms often rely on the measurement of two important aspects: relevance (selected elements should be important) and non-redundancy (duplicated content should be avoided). These two aspects are usually addressed by computing separate scores and deciding for the best candidates regarding some relevance redundancy trade-off. Summarization algorithms can be

* Corresponding author at: Lehrstuhl für Informatik 5, Universität Erlangen-Nürnberg, Martensstraße 3, 91058 Erlangen, Germany. Tel.: +49 (0)9131 8527879; fax: +49 (0)9131 303811.
E-mail addresses: korbinian.riedhammer@informatik.uni-erlangen.de (K. Riedhammer), benoit.favre@lium.univ-lemans.fr (B. Favre), dilek@icsi.berkeley.edu (D. Hakkani-Tür).

[1] As the presented algorithms can be applied in both text and speech summarization, we use "utterance" and "sentence" interchangeably.

0167-6393/$ - see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.specom.2010.06.002

categorized as *supervised* or *unsupervised*. A supervised system *learns* how to extract sentences given example documents and respective summaries. An unsupervised system generates a summary while only accessing the target document. Furthermore, the summarization problem can be specified as *single-document*, i.e., produce a summary for an independent document, or *multi-document*, i.e., produce a summary to represent a set of documents which usually cover a similar topic.

For this work, we focus on unsupervised methods. On the one hand, unsupervised methods are very enticing for meeting summarization as they do not depend on extensive manually annotated in-domain training data. They can thus be applied to any new observed data without (or only little) prior adjustments. On the other hand, we only compare unsupervised systems as it is rather unfair to compare unsupervised and supervised systems which are usually applied under different circumstances. If there is enough training data available for the required application, a supervised system may be the method of choice as long as training and test data are from the same domain. If, however, training data is not available, sparse or the test condition is unknown, unsupervised approaches should be considered. This is the case for our scenario as we are interested in a system that can summarize any kind of meeting without prior adjustment or retraining. Nonetheless, to give an idea of the performance of supervised systems, we include experiments with a classification baseline.

Some of the methods presented in this work are rooted in multi-document summarization. We do not use them for their ability to tackle the redundancy naturally occurring in a set of documents on the same topic, but rather to promote diversity in the generated summaries so that even minor topics discussed in a meeting are represented. Diversity is less of an issue in the supervised setup because sentences are represented according to a variety of orthogonal features (position, length, speaker role, cue words, ...) which each can lead to relevance. In the unsupervised setup, sentences with the same topical words get similar relevance assessments even if they are pronounced in very different contexts.

The most widely known algorithm for unsupervised summarization is maximum marginal relevance (MMR; Carbonell and Goldstein, 1998). This algorithm iteratively selects the sentence that is most relevant and least redundant to the previously selected ones. The greedy process can thus result in a suboptimal set of sentences as a better selection might be obtained by not choosing the most relevant sentence in the first place.

In this article, we are interested in inference models that seek a global selection of sentences according to relevance and redundancy criteria. Our contributions are as follow:

- We compare two approaches for global modeling in summarization: sentence-based scoring of relevance and redundancy, and sub-sentence based scoring with implicit redundancy.
  – For sentence-based scoring, we first propose an global formulation for MMR as an integer linear program (ILP). Such a formulation was not proposed before because of non-linearities in MMR. Then, we compare this formulation to the similar model by (McDonald, 2007) which relaxes the non-linearities to a linear function.
  – We outline a different approach to summarization which does not rely on sentence level assessment of redundancy and relevance. Instead, the quality of the summary is determined by the number of important concepts (sub-sentence units) covered. A selection of sentences satisfying this criterion is found again by solving an ILP. This approach is based on the ICSI text summarization system (Gillick et al., 2008; Gillick and Favre, 2009) and was modified for the meeting domain in (Gillick et al., 2009).
- While most MMR implementations rely on words and their frequency throughout the data, we could already show that using keyphrases instead of words to model relevancy leads to better performance for meeting summarization (Riedhammer et al., 2008a). In addition, keyphrases are used as concepts in the sub-sentence scoring approach. For this work, we refine keyphrase extraction and explore effects of pruning.
- We compare the complexity of the presented approaches and observe that sentence level models are less scalable than the concept level one.
- A comprehensive analysis of the summarization performance according to parameters, pruning and length constraints shows that the concept level model yields better properties than the others.

Throughout this work, the we use what we call “keyphrases”. Instead of extracting individual important words commonly known as “keywords”, we extract frequent noun phrases that match a certain pattern of determiners, adjectives and nouns.

This article is structured as follows. We begin with an overview of the related work in Section 2. In Section 4, we describe the two types of summarization models used for this work: sentence and concept based. For sentence-based summarization, we introduce a global formulation for the greedy MMR algorithm as an ILP and discuss how it relates to the formulation in (McDonald, 2007). For concept-based summarization, we present a model that gives credit to the presence of relevant keyphrases in the summary but penalizes them when they occur multiple times and discuss differences to similar approaches as found in (Filatova and Hatzivassiloglou, 2004; Takamura and Okumura, 2009). We conclude the model section with a description of how to extract the keyphrases which are the basis for both models. In Section 6, we describe the experiments we conducted to analyze the performance of the different approaches under fixed and varying constraints, compare greedy to optimal utterance selection and discuss two example summaries. We conclude with a discussion of the scalability of the methods and their flexibility towards practical use and, in a second step, abstractive summarization.

2. Related work

Speech summarization originated from the porting of methods developed for text summarization. It has been applied to various genres: broadcast news (Hori et al., 2002; Christensen et al., 2004; Zhang and Fung, 2007; Inoue et al., 2004; Maskey and Hirschberg, 2005; Mrozinski et al., 2005), lectures (Mrozinski et al., 2005; Furui et al., 2004), telephone dialogs (Zechner, 2002; Zhu and Penn, 2006) and meeting conversations (Murray et al., 2005a; Liu and Xie, 2008; Riedhammer et al., 2008b). Each genre brings different problems and is best summarized by different approaches. For example, while summarizing broadcast news is very similar to the summarization of textual documents, conversations are much less structured and involve the interaction between multiple speakers.

Approaches for speech summarization are mostly extractive and result in a selection of sentences from the input utterances. Some approaches also use sentence compression for removing superfluous words within sentences (Hori et al., 2002; Furui et al., 2004; Liu and Liu, 2009).

Sentence selection systems can be categorized as unsupervised or supervised. The former, which does not require training data, is represented by algorithms ported from the text community, such as variants of MMR (Murray et al., 2005a; Riedhammer et al., 2008a), graph based methods (Garg et al., 2009, in press, and concept-based methods (Filatova and Hatzivassiloglou, 2004; Riedhammer et al., 2008b; Takamura and Okumura, 2009).

Supervised approaches rely on a classifier, usually a support vector machine (Burges, 1998), to predict a binary class label for each input sentence indicating whether it should be included in the summary or not. Textual, structural and acoustic features have been developed for use in such approaches. Textual features include TFxIDF derivatives from the information retrieval community which assess the importance of a sentence according to the frequency of its words in the audio recording (Christensen et al., 2004). Sentence position and length, speaker role and dialog act type have been proved to be useful structural features (Murray et al., 2006). Fundamental frequency and energy contour, speaking rate, pauses, presence of disfluencies and repetitions have been used for characterizing relevant sentences (Maskey and Hirschberg, 2005; Zhu and Penn, 2006; Inoue et al., 2004; Xie et al., 2009b).

Evaluation of speech summarization is quite difficult because no gold-standard truth is available. Instead, multiple judges annotate sentences and write abstracts from which a metric, e.g., ROUGE (Lin, 2004), Pyramid (Nenkova and Passonneau, 2004), Basic Elements (BE; Hovy et al., 2006), is applied to evaluate the quality of the result (Hori and Furui, 2000; Murray et al., 2005b; Liu and Liu, 2008). For classification tasks, a weighted precision measure was introduced in (Murray and Renals, 2007). However, the fact that there might be two utterances with approximately the same wording but only one in the ground truth (thus awarding zero score if the other was extracted) leads to little adoption of this method in favor of the content oriented evaluations.

Baselines, such as the first sentences, a random selection, or the longest sentences can be used to calibrate results (Riedhammer et al., 2008a; Penn and Zhu, 2008). An alternative to automatic evaluation is to assess the usefulness of generated summaries on an information retrieval task (Murray et al., 2008), however this kind of evaluation involving humans is more expensive to perform.

3. Data

For the experiments described in this work, we used manual and ASR transcripts of the ICSI (Janin et al., 2003) and AMI (McCowan et al., 2005) meeting corpora.

The AMI meeting corpus consists of both scenario (i.e., the topic is given) and non-scenario meetings. For this work, we use a subset of 137 scenario meetings in which four participants play different roles in an imaginary company. They talk about the design and realization of a new kind of remote control. Though the topic was given, actions and speech are considered to be spontaneous as there was no specific script. All the meetings were transcribed and annotated with dialog act level relevance judgments and abstractive summaries, that is, human subjects summarized each meeting in their own words (about 300 words on average). There is one summary for each meeting. The AMI documentation provides a test set of 20 meetings, namely the series *ES2004, ES2014, IS1009, TS3003* and *TS3007*. Besides this subset, we also use the complete data set. Automatic transcripts were provided by the AMI ASR team (e.g., Renals et al., 2007), yielding a word error rate (WER) of about 36%.

For the ICSI meeting corpus, 75 regularly scheduled group meetings at the International Computer Science Institute at Berkeley were recorded, each lasting about 45 min. For this work, we use a subset of 57 meetings which have been transcribed and annotated with dialog acts and abstractive summaries (about 500 words on average). Following prior work on the ICSI corpus, we use a test set of six meetings: *Bed*{004, 009, 016}, *Bmr*{005, 019} and *Bro018*. For this subset, three human abstracts are available for each meeting. For the remaining ones, only one abstract is available. The speech recognition transcripts were provided by SRI International conversational telephone speech system (Zhu et al., 2005) and show a WER of about 37%.

4. Summarization models

In this section, we detail two models for extractive summarization based on sentence level and concept level scoring. For each of them, we present exact global inference algorithms in form of an ILP which are then solved using the open source ILP solver `glpsol` from the *GNU Linear Programming Kit*.[2]

4.1. Sentence based model

Most extractive summarization models rely on an assessment of the suitability of sentences for inclusion in a summary. Then, the most suitable sentences are selected and juxtaposed to form a summary. However, this approach can fail if sentences that convey the same information both have high scores leading to an inclusion of both sentences (e.g., in a classification approach). Hence, one needs to find a way of accounting for redundancy. This is generally implemented as a penalization of relevant sentences by a measure of their redundancy to the other sentences in the summary. Redundancy-penalized summaries tend to include more diverse information, which is important even in the single-document summarization setup.

The well-known MMR is a greedy algorithm that iteratively selects the most relevant sentence with respect to its similarity to the most similar sentence that was already selected for inclusion in the summary. Formally, the MMR score of sentence i can be expressed as

$$\mathrm{MMR}_i = \lambda \mathrm{Rel}_i - (1-\lambda) \max_{j \in S} \mathrm{Red}_{ij}, \tag{1}$$

where Rel_i is the relevance score of sentence i and Red_{ij} is the redundancy penalty for having both sentence i and j in the summary S. The algorithm terminates when a summary length constraint is reached. The definition of relevance and redundancy measures that discriminate well between sentences will be described in Section 5.

The greedy nature of this algorithm implies that a sentence, once selected, is not reconsidered in favor of other sentences. Therefore, it is likely that the final selection is suboptimal. For example, two shorter sentences could be selected in place of a longer one in order to provide more information within the length constraint. This problem can be addressed by considering a global objective function

$$\text{Maximize}: \quad \sum_i [\lambda \mathrm{Rel}_i s_i - (1-\lambda) \max_j \mathrm{Red}_{ij} s_i s_j] \tag{2}$$

$$\text{Subject to}: \quad \sum_i l_i s_i \leqslant L. \tag{3}$$

Here, s_i represents a binary indicator of the presence of sentence i in the summary, l_i is the length of sentence i and L is the length limit for the whole summary.

Finding an optimal assignment of s_i, $\forall i$ for the MMR's global formulation requires solving a 0–1 quadratic problem which includes a $\max(\cdot)$, making it non-linear. An approximate solution can be found by various optimization techniques such as Monte-Carlo search or genetic programming. Nevertheless, (McDonald, 2007) proposed to change the global MMR formulation in order to make it a linear problem and introduced additional constraints in order to obtain a solvable ILP

$$\text{Maximize}: \quad \sum_i \left[\lambda \mathrm{Rel}_i s_i - (1-\lambda) \sum_{j \neq i} \mathrm{Red}_{ij} s_{ij}\right] \tag{4}$$

$$\begin{aligned}
\text{Subject to}: \quad & s_{ij} \leqslant s_i \quad \forall i,j, && (5)\\
& s_{ij} \leqslant s_j \quad \forall i,j && (6)\\
& s_i + s_j \leqslant 1 + s_{ij} \quad \forall i,j && (7)\\
& \sum_i l_i s_i \leqslant L. && (8)
\end{aligned}$$

The constraints in this formulation assure that s_{ij}, a binary indicator of presence of the sentence pair i and j in the summary, will be 1 if and only if both s_i and s_j equal 1. The $\max(\cdot)$ in the redundancy term is replaced by a sum which roughly corresponds to penalizing a sentence according to its average redundancy to the other sentences in the summary. McDonald's formulation was the first to be proposed for global inference in summarization. At this point, it is appealing to express the global MMR using an ILP in the same way McDonald did in order to reach optimal solutions. This can be achieved by converting the inner working of the $\max(\cdot)$ to ILP constraints

$$\text{Maximize}: \quad \sum_i \left[\lambda \mathrm{Rel}_i s_i - (1-\lambda) \sum_{j \neq i} \mathrm{Red}_{ij} m_{ij}\right] \tag{9}$$

$$\begin{aligned}
\text{Subject to}: \quad & \sum_j m_{ij} = s_i \quad \forall i, && (10)\\
& m_{ik} \geqslant s_k - (1 - s_i) - \sum_{j:\mathrm{Red}_{ij} \geqslant \mathrm{Red}_{ik}} s_j \quad \forall i \neq k, && (11)\\
& m_{ij} \leqslant s_i \quad \forall i, && (12)\\
& m_{ij} \leqslant s_j \quad \forall j, && (13)\\
& \sum_i l_i s_i \leqslant L. && (14)
\end{aligned}$$

Here, we introduce m_{ik} as a binary indicator for Red_{ik} to be the max among the $\mathrm{Red}_{i(*)}$ for all sentences included in the summary. The idea is to explicitly compute which sentence of the summary is most redundant to which other sentence of the summary. For each sentence, the other sentences are ordered by their respective redundancy. Then, from this sorted list, we only look at the selected sentences. Once a sentence is selected it requires exactly one other selected sentence to be considered the most redundant one (Eq. (10)). For any $m_{ik} = 1$, i.e., sentence k is maximum redundant regarding sentence i, both sentences i and k need to be in the summary (Eqs. (12) and (13)) and no sentence with a higher redundancy to i can be selected (Eq. (11)). This formulation has more constraints than the original formulation by McDonald, however, it gets rid of the linear approximation and is therefore an optimal solution to the MMR problem.

[2] http://www.gnu.org/software/glpk/.

Please cite this article in press as: Riedhammer, K., et al. Long story short – Global unsupervised models for keyphrase based meeting summarization. Speech Comm. (2010), doi:10.1016/j.specom.2010.06.002

4.2. Concept-based model

In the previously presented models, sentence level redundancy assessment is limited to pairs of sentences. Redundancy introduced in a summary by groups of more than two sentences is out of the scope of these models. Fig. 1 draws an example where a set of two sentences completely entail a third sentence, a fact that does not prevail if redundancy is computed pairwise.

In (Gillick et al.,2008, 2009), we proposed a more natural way of estimating both relevance and redundancy in a global inference framework for summarization based on integer linear programming. Concept based summarization assumes that the information can be expressed in term of concepts. Concepts can be facts, events, or information units that characterize relevant content, such as the keyphrases that will be defined in Section 5. Each concept appearing in the summary is given credit only once, in order to penalize the use of the same information in multiple sentences. This approach goes beyond pairs of sentences to tackle both relevancy and redundancy in the whole summary.

The idea of concepts has been around for some time especially in the text summarization community. Evaluation measures for summarization performance like ROUGE (Lin, 2004) or Pyramid (Nenkova and Passonneau, 2004) and later developments like Basic Elements (Hovy et al., 2006) score summaries based on an overlap of n-grams (ROUGE), summary content units (manually annotated parts in the target text; Pyramid) or dependency parsing relations (Basic Elements).

Formally, let c_i denote the presence of concept i in the summary and s_j denote the presence of sentence j in the summary. Each concept can appear in multiple sentences and sentences can contain multiple concepts. The occurrence of concept i in sentence j is denoted by the binary variable o_{ij}. The score of a summary is expressed as the sum of the positive weights w_i of the concepts present in the summary. The length of the summary is limited by a constant L over the sum of the length l_j of its sentences. Finding the summary that has the maximum score can again be expressed as an ILP

$$\text{Maximize}: \quad \sum_i w_i c_i \tag{15}$$

$$\text{Subject to}: \quad \sum_j s_j l_j \leqslant L, \tag{16}$$

$$s_j o_{ij} \leqslant c_i \quad \forall i, j, \tag{17}$$

$$\sum_j s_j o_{ij} \geqslant c_i \quad \forall i. \tag{18}$$

(1)	The device should be white.
(2)	The device should be round.
(3)	The device should be round and white.

Fig. 1. Redundancy in a group. The pairwise redundancy scores will not indicate that (1) with (2) conveys the same meaning as (3).

In this ILP, the objective function is maximized over the weighted sum of the concepts present in the summary given the length constraint. Consistency constraints ensure that if a sentence is selected, all concepts it contains are also selected and if a concept is selected, at least one sentence that contains it is selected. In detail, Eq. (18) ensures that if a concept i is in the summary, then there is at least one summary sentence covering it. Eq. (17) assures that every concept i that appears in the summary ($s_j o_{ij} = 1$) is actually incorporated in the objective function by enforcing $s_j o_{ij} = 1 \Rightarrow c_i = 1$.

This model extends prior related work. Filatova and Hatzivassiloglou (2004) were probably the first to use units similar to our concepts. They call them events, and find a selection of sentences that maximize event coverage using an adaptive greedy algorithm. Independently and from our work, Takamura and Okumura (2009) introduced an ILP formulation very similar to our previously published text summarization system (Gillick et al., 2008) for what they call the Maximum Coverage Problem with Knapsack Constraints (MCKP). In fact, their formulation is equivalent to ours without constraints from Eq. (17). Without these additional constraints, the objective function can be skewed, i.e., there might be concepts in the summary which do not contribute to the score. This might also be the reason why in their comparison of different strategies the exact solution (obtained by branch-and-bound) was not necessarily superior to approximations like the greedy solution or stack decoding.

4.3. Supervised baseline

To give an idea how the previously described unsupervised methods perform compared to a supervised system, we briefly introduce a supervised baseline. For each input sentence, a set of features is extracted and fed to a classifier in order to predict binary relevance labels as annotated in the AMI and ICSI data.

For this work, we consider the following features for each utterance which are extracted from the manual transcriptions and annotations.

- *Duration* of the utterance in seconds.
- *Position* of the utterance in terms of the start time relative to the meeting duration.
- *Speaker dominance* in terms of how much the speaker spoke compared to the others.
- *Speaker role*, e.g., professor (ICSI data) or product manager (AMI data).
- *Word n-grams*. For each word n-gram in the corpus, the value is 1 if it appears in the utterance or 0 otherwise.
- *Dialog act*, i.e., the type of utterance, e.g., question or answer.

These features, among others, have successfully been used for supervised meeting summarization (e.g., Xie et al., 2009b). For speaker role and dialog acts, we used manual annotation of these features in the corpus.

For generating relevance predictions, we rely on an Adaboost variant (Boostexter[3]; Schapire and Singer, 2000) that iteratively selects the best features while re-weighting examples in order to focus on more difficult ones (it often gives as good predictions as SVMs). Sentences with the highest relevance prediction are selected until the length constraint is fulfilled.

5. Relevance, redundancy and concepts

Though the previous section provides theoretical models required to build the summarization systems, the question of how to measure relevance and redundancy and how to find the concepts remains open. In text summarization, relevance is usually defined by a (user generated) query. The relevance score of a candidate sentence is then determined by an overlap measure with that query; redundancy is modeled in a similar way. If no query is provided, an artificial query is generated to represent the overall gist of the text.

In (Filatova and Hatzivassiloglou, 2004), the authors extracted "atomic events" from written language to use as concepts which are basically pairs of named entities ("relations") and the words in-between ("connectors"). The connectors are further reduced to content verbs or action nouns using an external information source (in their case WordNet). The concepts are weighted by their normalized relation and connector frequency. In (Takamura and Okumura, 2009), the authors use words and related weights obtained by either an unsupervised "interpolated weight" computed from the generative word probability in the entire document and that in the beginning part (100 words), or a "trained weight" which is learned using logistic regression on training instances whether or not a word appears in the training summary or not.

Unfortunately, spontaneous multi-party speech strongly differs from text or even structured speech (e.g., broadcast news read from a teleprompter). The presence of disfluencies, restarted sentences, repetitions, filled pauses (e.g., "ahm", "hm"), idioms and speaker-specific sayings (e.g., "To my mind, (*what the speaker actually wanted to say*), right?") makes it hard to compute reliable statistics about the importance of the individual words spoken.

However, spontaneous multi-party speech suggests the use of a fairly simple heuristic. In contrast to text, where sometimes different words are used to express the same meaning, people tend to use the same phrases as other discourse participants (and also stick to that phrase throughout the whole conversation) in order to find a common ground for their communication. To be more specific, things of interest to all speakers will be called the same name by all speakers. We call these keyphrases.

Using keyphrases to model relevance, redundancy and concepts has already shown to outperform previous word based models (Riedhammer et al., 2008a; Gillick et al., 2009) and also provides a common ground for a fair comparison of sentence and concept-based summarization models.

5.1. Keyphrase extraction

Though keyphrases can also be extracted using a classification system (e.g., Liu et al., 2008) we believe that unsupervised methods are the method of choice as training data is rare and highly domain specific. We refined the extraction procedure from (Riedhammer et al., 2008a) as follows:

1. Apply part-of-speech (PoS) tagging.
2. Extract all word n-grams g_j if the respective PoS tag n-gram matches a regular expression of determiners, adjectives and nouns.[4] This step allows to catch complex noun phrases like "trained network of individual nodes" without requiring a proper parse tree.
3. Noise reduction: Remove unique and enclosed n-grams (e.g., "manager" if it occurs as many times as the phrase "program manager").
4. Re-weight n-grams in order to emphasize the occurrence of longer keyphrases: $w_j = \text{frequency}(g_j) \cdot (n+1)$, $n > 1$ where w_j is the final weight and n is the n-gram length. That means that the longer a repeated keyphrase is, the more likely it is that the repetition was on purpose, thus of interest.

The re-weighting in the last step is still rather biased towards shorter keyphrases due to its linear design. A study on English and Chinese text data showed that bi-gram frequencies are about an order of magnitude larger than 5-gram frequencies (Ha et al., 2002). Also, results on the keyphrase extraction given later in Section 6 indicate a rather exponential decay of noun phrase frequencies with increasing length. However, we first try a rather strong approximation in form of a linear weighting to accommodate for the fact in general and having in mind that Zipf's law (and any other statistic) usually only hold for very large data which is not the case for the present work. For future work on larger data, modifying the weighting is definitely of interest.

Recent work on unsupervised keyphrase extraction integrates TFxIDF and graph based models (Liu et al., 2009). However, the focus of our work is to compare sentence and concept-based summarization. Also, it remains unclear if the keyphrases extracted in (Liu et al., 2009) are of better quality, as the authors did not provide summarization results and we could not compare our approach within their evaluation setup.

[3] We use the icsiboost implementation, available at http://code.google.com/p/icsiboost.

[4] $JJ^*(NN|NNS|FW|CD)^+((DT|IN)^+JJ^*(NN|NNS|FW|CD)^+)^*$. A list of the tags and their meaning can be found for example in (Santorini, 1990).

5.2. Relevance and redundancy

For the utterance based model, relevance and redundancy are defined as in (Riedhammer et al., 2008a). The former is a sum over the occurring keyphrases (binary indicator function "occ(g_j, i)" returns 1 if g_j occurs at least once in sentence i) while the latter is a normalized non-stopword word overlap. The stopword list contains about 500 words and includes pronouns, articles, particles and other frequent function words in order to not distort the redundancy score[5]

$$\mathrm{Rel}_i = \sum_j \mathrm{occ}(g_j, i) \cdot w_j;$$

$$\mathrm{Red}_{ij} = \frac{\mathrm{words}(i) \cap \mathrm{words}(j)}{\max(\mathrm{words}(i), \mathrm{words}(j))}. \tag{19}$$

We understand that the chosen relevance and redundancy scores are rather simple. However, it is important to base all our models on the same ground in order to get a fair comparison. For the redundancy score, prior experiments have shown that a normalized word overlap is sufficient when using MMR and keyphrases for meeting summarization. The fact that two utterances containing the same concepts but having different lengths will result in the same redundancy score (due to the maximum operator) is compensated in the optimization process which inherently favors shorter sentences in presence of same relevance and redundancy.

5.3. Concepts

For the concept-based model, each keyphrase is handled as an individual concept. A concept is assigned to an utterance if it occurs at least once. In case of enclosing keyphrases (e.g., "manager" in presence of "system manager") one can decide to assign only the longest matching one instead, thus ignoring the keyphrases with less context.

6. Experiments

From the theory described in Sections 4 and 5, we build several summarizers to analyze and compare the performance of utterance and concept-based systems:

- *mmr/greedy*: The original iterative (greedy) MMR using keyphrase similarity as relevance and word overlap as redundancy measure.
- *mmr/ilp*: The proposed ILP for a global formulation of MMR using the same relevance and redundancy scores as above.
- *mcd/ilp*: McDonald's ILP formulation for global inference (McDonald, 2007) for comparison.
- *concepts/grd and concepts/ilp*: Global formulation with concept-based summarization using keyphrases as concepts, greedy (*grd*) and optimal (*ilp*) solution respectively. In case of enclosing keyphrases within an utterance, only the longest matching keyphrase is assigned. For the greedy solution, the utterances with the highest keyphrase weight were selected in an iterative manner.Furthermore, we build a classification system learned on the training subsets of the data to give an idea about performance of supervised systems on the same setup.
- *supervised*: The supervised baseline using both textual and higher level speech features.

It is difficult to compare to supervised systems found in the literature because they are generally scored against *extracts* (the concatenation of all relevant utterances) while we compute performance against human-written *abstracts*.

The experiments are divided into three parts, and performed using manual transcripts unless stated otherwise. First, we analyze the performance of the different systems in an evaluation setup which is fixed in terms of length and parameters to ensure a fair comparison. Second, we analyze how system performance vary if these constraints are changed in order to see if one system always outperforms another. Finally, we investigate the effect of the tunable λ parameters, and the way of assigning keyphrases to utterances.

We compare the automatic summaries to the human abstracts using ROUGE-1, 2 and SU4 which basically determine n-gram overlap between reference (human) and system summaries, ignoring stopwords as built into the ROUGE package (Lin, 2004). We consider ROUGE-1 to be the most fair measure when comparing spontaneous speech extracts to written language, as higher n-gram overlap is rather unlikely to be found given how these two different data look like.

6.1. Keyphrase extraction

For keyphrase extraction, we use a part-of-speech tagger based on (Thede and Harper, 1999; Huang et al., 2007). The models trained on English broadcast news were provided by the referenced authors. To give an example, after stopword removal, the top five keyphrases for the AMI meeting *ES2004c* are "remote control", "button", "design", "voice recognition" and "rubber", which makes good sense recalling the topic of this meeting. Table 1 shows how many n-gram keyphrases could be extracted from the data. It is interesting to see that the number of keyphrases drops exponentially as n gets larger. Summary examples for meeting *ES2004c* will be displayed at the end of this section.

6.2. Fixed lengths

For the first part, we chose to generate summaries of 300 words for the AMI meetings and of 500 words for the ICSI

[5] Computing only keyphrase overlap is not advisable as this leads to many similar or equal scores which is not desirable for the later optimization process

meetings. These fixed lengths were chosen to match the average length of the human abstracts and following the idea that a user might prefer summaries of fixed (short) length instead of a variable length (think of typical "minutes" or executive statements). Additionally, we set the relevance parameter for MMR variants to $\lambda = 0.9$ based on findings in Section 6.4.1.

As both approaches come down to an optimization problem, we provide greedy and global solutions, as long as they were computable in reasonable time: For *mmr/ilp* and *mcd/ilp*, we reduced the number of candidate utterances to the top 50 in terms of the sum of the keyphrase weights, in order to obtain a more feasible problem. As runtimes turned out to be rather long, we additionally restricted computation time to a maximum of 60 min, deciding for the best current solution at that time limit (we will give further comments on runtime later this section). Note that this is an approximation in terms of *computational power* instead of an approximation of *modeling redundancy* as in MMR. One should have in mind that the obtained solution might have been better if more compute power were available.

For completeness and better comparison, we add results of the classification baseline *supervised* and systems used in previous work: *baseline1* (longest utterances first), *baseline2* (greedy MMR using a term frequency based centroid term vector of the meeting, and cosine similarity) and *max-r* (ROUGE-1 recall oracle), as described for example in (Riedhammer et al., 2008b).

6.2.1. Results in comparison

Table 2 shows the results for the complete and test sets using manual transcriptions. For AMI data, a clear ranking can be read. From *baseline1*, performance significantly increases for *baseline2* to *concepts/ilp* and the oracle *max-r*, for both complete and test sets. A similar observation holds for the ICSI data, although *baseline2* performs worse than *baseline1* and *mmr/ilp* is outperformed by its original greedy formulation.

The *supervised* system was only evaluated on the test set as the rest of the data is used for training. Performance is similar to the concept-based systems for all evaluation metrics which suggests that unsupervised systems can produce competitive results. As supervised approaches are not the focus of this work, they will not be considered any further.

ROUGE-2 and ROUGE-SU4 results in Table 2 are lower than ROUGE-1 which is expected as the overlap in *n*-grams between the reference transcripts and the handwritten abstracts is relatively low because of intrinsic differences in style. The consequence is a smaller spread of the scores and a less clear ranking of the systems even though the trend is respected. For the remainder of the analyses, we will only display ROUGE-1 scores for clarity and conciseness. Note that none of the systems is particularly designed to get better scores on ROUGE-1 rather than on the other two metrics.

6.2.2. Significance and runtime analysis

The significance chart given in Table 3 confirms the above system ranking, however, two aspects are worth further analysis.

The lack of significance of the performance increase of the global sentence-level systems *mmr/ilp* and *mcd/ilp* compared to the greedy *mmr* needs to be explained. In theory, the formulas should lead to a better result than the original, greedy formulation, assuming good relevance and redundancy measures. In practice however, the global systems seem to be hurt by the complexity of the problem they have to solve: The number of constraints increases by $O(n^3)$, where n is the number of utterances. Thus, the more utterances, the more time is potentially needed to some the optimization problem. This was also discussed in (McDonald, 2007) where the number of sentences had to be reduced to 100 for computational feasibility. As mentioned in the beginning of this section, we limited the computation time to 60 min per meeting and reduced the number of utterances to the 50 highest scoring ones according to their keyphrase weight in order to retrieve a (possibly suboptimal) result in reasonable time. That implies on the one hand that we might have stripped out potential good candidates as well as we possibly stop the optimization in a non-optimal state. If the optimization was stopped prematurely, the current best solution is used. It is possible but not necessary that the optimal solution for the given input differs from the current solution.

In fact, for *mmr/ilp*, only 19 (2) out of 137 (57) of the AMI (ICSI) summaries did not exceed the time constraint. Similarly but better, for *mcd/ilp*, 51 (7) out of 137 (57) optimization problems finished in time. Note that in these cases, the solver reported solutions within 1–2% (in value) of the estimated maximum objective function, which validates results as close to actual optimal solutions.

Looking at the results for the ICSI test set, *mmr/ilp* reveals a (not significantly) weaker performance than the original formulation. A possible explanation for this might be found in the implementation of the ILP solver: The one used for this work (`glpsol`) first determines a floating

Table 1
Number of keyphrases for AMI and ICSI meetings. The numbers in parentheses indicate the average number per meeting.

n	AMI (avg.)	ICSI (avg.)	Total
1	13366 (100)	11036 (187)	24402
2	4191 (31)	3866 (66)	8057
3	596 (4)	641 (11)	1237
4	142 (1)	197 (3)	339
5	55	47	102
6	11	14	25
7	0	4	4
8	0	1	1
9	0	3	3
Avg.	134	268	–

Table 2
ROUGE-1, 2 and SU4 F scores on the complete and test sets using manual transcriptions. For systems *mmr/ilp* and *mcd/ilp*, the number of utterances was reduced to 50 in order to allow a feasible optimization. Summary length is 300 words for AMI and 500 words for ICSI meetings.

	ROUGE-1				ROUGE-2				ROUGE-SU4			
	AMI		ICSI		AMI		ICSI		AMI		ICSI	
	All	Test	All	Test	All	Test	All	Test	All	Test	All	Test
baseline1	.19	.17	.17	.16	.03	.02	.02	.02	.05	.04	.04	.04
baseline2	.21	.21	.17	.15	.04	.04	.02	.02	.07	.07	.04	.04
mmr/greedy	.23	.22	.18	.17	.04	.04	.03	.03	.08	.07	.05	.04
mmr/ilp	.24	.23	.19	.16	.05	.05	.03	.03	.08	.07	.05	.04
mcd/ilp	.25	.25	.20	.18	.05	.04	.03	.02	.08	.07	.05	.04
concepts/grd	.26	.25	.22	.21	.05	.04	.02	.03	.09	.09	.06	.05
concepts/ilp	**.28**	**.29**	**.23**	**.22**	**.06**	**.05**	**.03**	**.03**	**.08**	**.09**	**.06**	**.06**
supervised	–	.25	–	.20	–	.04	–	.03	–	.07	–	.05
max-r	.46	.47	.41	.33	.11	.11	.06	.05	.15	.15	.11	.09

Table 3
Table of significant improvements for AMI/ICSI manual transcripts; read "row system significantly outperforms column system" (setup as in Table 2).

	baseline1	*baseline2*	*mmr/greedy*	*mmr/ilp*	*mcd/ilp*	*concepts/greedy*
baseline2	✓/–					
mmr/greedy	✓/–	✓/–				
mmr/ilp	✓/✓	✓/✓	–/–			
mcd/ilp	✓/✓	✓/✓	✓/✓	✓/–		
concepts/grd	✓/✓	✓/✓	✓/✓	✓/✓	–/✓	
concepts/ilp	✓/✓	✓/✓	✓/✓	✓/✓	✓/✓	✓/✓

point solution and then tries to find the best fitting integer solution as a second step which most likely differs from the greedy path.

A closer look at the similarity and redundancy values revealed the difficulty for the solver. Once the current solution contains all the utterances with high relevance, the remaining ones all have very similar or even equal relevance and redundancy scores. This leads to many selections with the same objective function value which need to be enumerated by the solver.

Unfortunately, there was no matching subset for the uncompleted optimizations of the utterance based ILP that would have allowed a closer look at the problem. Also we chose not to increase the amount of computation time as we are interested in a scalable and fast method – a system requiring many hours to produce a summary does not seem acceptable by users.

Other than computational concerns, the fact that greedy solutions are not worse than global ones can be imputed to the relevance and redundancy metrics that would be valuable in the greedy case (as shown in previous work) but not adapted to the global case. Moreover, humans do not compute similarity between sentences for selecting them in a summary, they devise the importance of facts that they contain, which is the motivation of our other global model.

The concept-based summarizer using keyphrases and ILP for optimization significantly outperforms all utterance based systems on all evaluation scenarios. This confirms previous results using a different, variable length based evaluation setup, as for example in (Gillick et al., 2009; Riedhammer et al., 2008a).

Additionally, the concept-based system shows better runtime and complexity properties. While the greedy solutions are the fastest (only a few milliseconds on a reasonably fast machine), the *concepts/ilp* system runs almost as fast while the complexity is mainly controlled by the number of keyphrases instead of the number of utterances. This is especially important for interactive systems as described for example in (Riedhammer et al., 2008a; Mieskes et al., 2007), for which a fast responding summarization algorithm is required to give the user immediate feedback. Pruning keyphrases is intuitively less destructive than pruning sentences as the former reduces the possibility for a sentence to be included rather then excluding it completely.

Both performance and runtime advantage match our findings in text summarization where the concept-based ILP system was top ranked in TAC'08 and TAC'09, at a runtime of about one second per summary with approximately 1000 sentences and 1000 concepts per instance. A comprehensive comparison of this model against McDonald's in term of scalability can be found in (Gillick and Favre, 2009).

Another interesting observation is that the oracle *max-r* is better for AMI data than for ICSI data, especially on the test set. This is due to the fact that there is only a single human reference summary for each AMI meeting but the ICSI test set provides three reference summaries for each meeting, making it harder to find a summary that matches all at the same time.

6.2.3. Results on ASR

To check the consistency of the above ranking in noisy conditions, we conducted the same experiments using

Please cite this article in press as: Riedhammer, K., et al. Long story short – Global unsupervised models for keyphrase based meeting summarization. Speech Comm. (2010), doi:10.1016/j.specom.2010.06.002

ASR transcripts for all algorithms, including keyphrase extraction. Note that the AMI meeting *IS1003b* is skipped due to a missing ASR transcript. As it is not part of the test set and there are 137 meetings in total, comparing the numbers to the ones given in Table 2 should be fair.

Table 4 gives an overview of the results. Beside a few exceptions due to rounding the numbers, the overall trend of Table 2 is confirmed. For ICSI data, *concepts/ilp* still significantly outperforms all other systems. For AMI data, both greedy and optimal solution to the concept-based approach significantly outperform the utterance based ones. However, the difference between the greedy and optimal solution is not significant anymore. This confirms observations in prior work that using ASR instead of manual transcription reduces performance, but does not affect the ranking of algorithms. That is, the loss of performance is directly related to the quality of the ASR output but not to the system design.

Interestingly, the performance loss is higher for the ICSI setup which can be best seen when comparing oracle *max-r* scores. The ROUGE-1 *F* score is only reduced from .46 (.47) to .44 (.45) for the AMI (test) set while values drop from .41 (.33) to .36 (.39) for the ICSI (test) set. The fact that this was observed for the oracle as well as for all other systems suggests that words responsible for good ROUGE scores are more affected than others by recognition errors given the more spontaneous ICSI data.

6.3. Variable lengths

The second part of the experiments is to analyze how the different systems behave under varying constraints. In Figs. 2 and 3, we show performance charts of the systems for different length constraints from 200 to 500 words, with a step size of 50. The *max-r* system is left out as it is off the chart for the given scale. With one (not significant) exception, the systems keep the ranking shown in Table 2, regardless of summary length. Given the same keyphrases and available utterances, the concept-based systems outperform the utterance based ones (compare contours *mcd/ilp* and *concepts/ilp* in Figs. 2 and 3) on all length constraints.

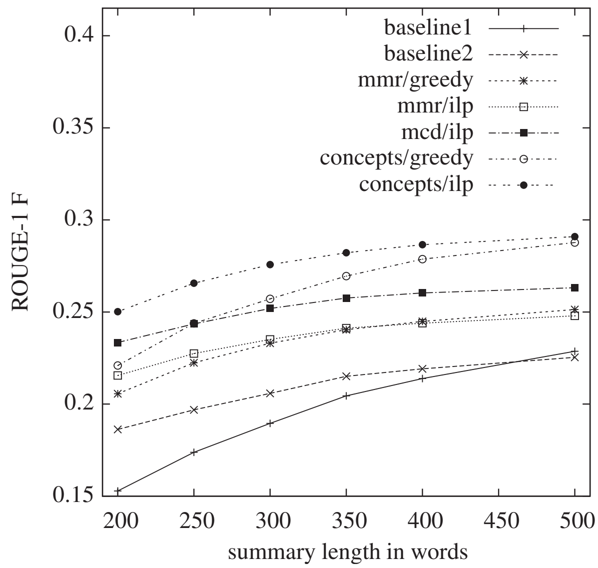

Fig. 2. Performance chart using all AMI meetings using manual transcripts; *max-r* is always above .41 and thus omitted from the chart.

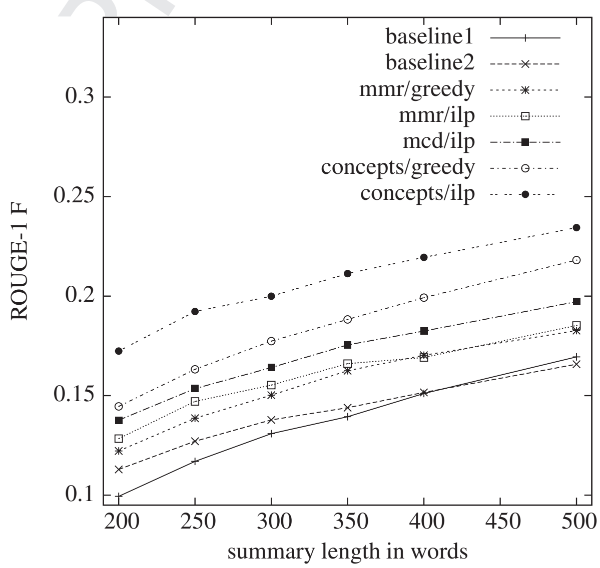

Fig. 3. Performance chart using all ICSI meetings using manual transcripts; *max-r* is always above .30 and thus omitted from the chart.

Table 4
ROUGE-1, 2 and SU4 F scores on the complete and test sets using ASR transcripts. For systems *mmr/ilp* and mcd/*ilp*, the number of utterances was reduced to 50 in order to allow a feasible optimization. Summary length is 300 words for AMI and 500 words for ICSI meetings.

	ROUGE-1				ROUGE-2				ROUGE-SU4			
	AMI		ICSI		AMI		ICSI		AMI		ICSI	
	All	Test	All	Test	All	Test	All	Test	All	Test	All	Test
baseline1	.21	.19	.10	.10	.03	.02	.01	.01	.06	.05	.02	.02
baseline2	.22	.22	.10	.09	.04	.04	.01	.01	.07	.07	.02	.02
mmr/greedy	.24	.24	.10	.10	.05	.04	.01	.01	.08	.07	.02	.02
mmr-ilp/kp	.24	.24	.10	.09	.04	.04	.01	.01	.08	.07	.02	.02
mcd-ilp/kp	.25	.25	.11	.11	.05	.04	.01	.01	.08	.07	.02	.02
concepts/grd	.27	.29	.14	.13	.05	.05	.01	.01	.08	.08	.03	.03
concepts/ilp	**.28**	**.30**	**.15**	**.16**	**.05**	**.05**	**.01**	**.01**	**.08**	**.08**	**.03**	**.03**
supervised	–	.25	–	.21	–	.04	–	.03	–	.07	–	.05
max-r	.44	.45	.36	.29	.09	.10	.03	.03	.14	.14	.09	.06

However, some of the utterance based ILP did not finish in the given time limit, as in the previous experiment.

6.4. Parameter tuning

6.4.1. Relevance parameter

For MMR variants, the relevance parameter λ has to be set either manually, or learned on some training set. To see whether or not our experiments where biased by choosing a fixed λ, we sample different values for λ and evaluate on the two length previously used to compute the results given in Table 2. Figs. 4 and 5 show the performance charts for AMI and ICSI data (complete sets). A higher lambda means more weight to relevance and less to redundancy, but it also embeds the scale of the two factors and should not be interpreted directly as evidence of redundancy of the data (remember that relevance is computed against the whole meeting). In our case, $\lambda = 0.9$ seems to be a reasonable choice for all the benchmarked algorithms (except for *mmr/ilp* on the ICSI data which peaks at 0.8), and note that $\lambda = 1$ is worse, emphasizing the importance of considering redundancy even though a single meeting is not likely to be redundant. This effect is probably due to the additional diversity of the content put in the summary when a topic dominates the meeting and skews relevance.

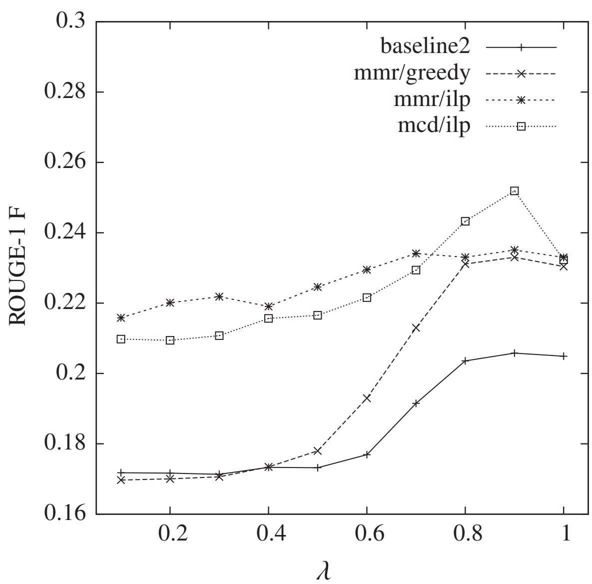

Fig. 4. Effect of the relevance parameter λ on the summarization score (300w, all AMI meetings, manual transcription).

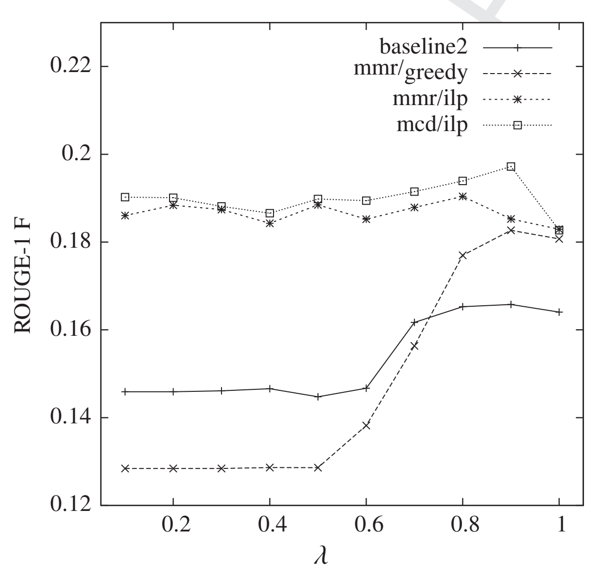

Fig. 5. Effect of the relevance parameter λ on the summarization score (500w, all ICSI meetings, manual transcription).

Further experiments using the fast *mmr/greedy* showed that this also holds for varying summary lengths. Interestingly, the ILP formulations are less sensitive to λ than the greedy variants. This indicates that the key to a good greedy solution is the proper selection of the relevance parameter. Also, at lower values of λ, the *mmr/ilp* system outperforms the less strict *mcd/ilp* on the AMI data set.

6.4.2. Keyphrase assignment

For the concept-based systems using keyphrases, we explored two parameters. The first parameter is to prune either the number of extractable utterances or the number of assignable keyphrases. For the first, we reduced the number of utterances to the top 50 in terms of the sum of the keyphrase weights as it was done for the utterance based systems. For the latter, we limited the number of keyphrases to the top 25 in terms of weight.

Second, when identifying concepts in an utterance, one can either account for all keyphrases, i.e., including redundant ones like “manager” in presence of “project manager”, or just account for the longest match, i.e., drop “manager” in presence of “project manager”.

As shown in Fig. 6 and 7, regardless of the summary length, dropping redundant keyphrases leads to the best results. Intuitively, pruning decreases summarization scores. The performance of the systems with reduced number of keyphrases stays at the same level for longer summary lengths as there are only a little number of utterances available for selection due to the small number of keyphrases.

6.5. Example summaries

Below is an example summary (about 300w) for the AMI meeting *ES2004c* generated by a human annotator and by the systems *mmr/ilp* and *concepts/ilp*. The automatic summaries are based on the manual transcriptions and the extracted utterances are ordered as they appear in the meeting. The contributing keyphrases are highlighted and their weight is shown in parentheses. Utterances occurring in both system summaries are typeset as italic.

It can be observed that the MMR based system favors longer sentences due to the implemented relevance scoring. The probably most interesting fact is, that the *mmr/ilp* summary covers only 46 unique keyphrases with a combined

Please cite this article in press as: Riedhammer, K., et al. Long story short – Global unsupervised models for keyphrase based meeting summarization. Speech Comm. (2010), doi:10.1016/j.specom.2010.06.002

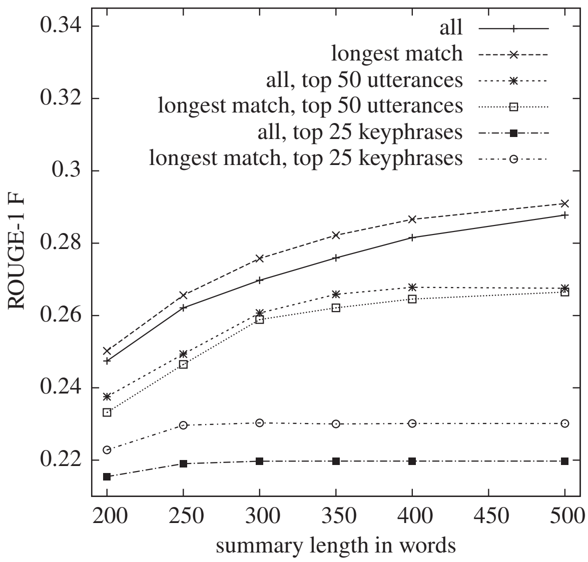

Fig. 6. Effect of pruning on summarization scores using *concepts/ilp* and all AMI meetings (manual transcripts).

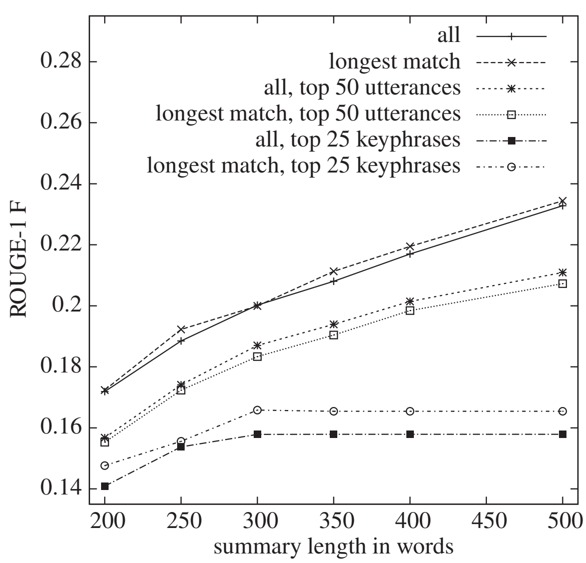

Fig. 7. Effect of pruning on summarization scores using *concepts/ilp* and all ICSI meetings (manual transcripts).

weight of 435 but the *concept/ilp* summary covers 88 unique keyphrases with a combined weight of 778, almost twice as much. However, the concept-based system tends to include shorter, possibly ill-formed or aborted sentences to yield a larger concept coverage which will be further addressed in the discussion. The human summary shows 40 unique keyphrases with a score of 302 and shows some redundancy due to the way the human subjects were instructed to design the abstract.

The summaries reveal some incorrectly extracted keyphrases like **thing**, **something** or **kind** which correspond to speaker idioms and represent less valuable content. Also, the extracts do not match the style of the abstracts, suggesting to work on spoken discourse reformulation.

6.5.1. Human

The project manager reviewed the decisions from the previous **meeting** (**7**). The **marketing** (**4**) expert made a presentation on **trend** (**5**) watching, including trends in **user** (**3**) requirements and trends in **fashion** (**5**). The industrial designer presented all the components of the **device** (**4**) and announced that several of the features already discussed would not be available. He suggested substituting a kinetic **battery** (**18**) for the rechargeable batteries and using a combination of **rubber** (**20**) and **plastic** (**8**) for the materials. The **user (3) interface** (**4**) designer presented his main **interface (4) design** (**22**), which included buttons for the most frequently used features and a **graphic_user_interface** (**8**) on the **lcd_screen** (**12**) for other functions, to keep frequently used features easy to use. He announced that speech **recognition** (**8**) was still an **option** (**8**) to consider, depending on price. The project manager then began a discussion to decide what was going into the final **design** (**22**). It was decided that a kinetic **battery** (**18**) would be used in place of a rechargeable **battery** (**18**), that the **remote** (**5**) will **feature** (**10**) an **lcd_screen** (**12**) and **rubber (20) casing** (**3**) and **rubber** (**20**) buttons, and that interchangeable **rubber** (**20**) covers in **fruit** (**7**) colors will be available. Speech **recognition** (**8**) may be included if it is not too costly. It was decided that the **remote** (**5**) would **feature** (**10**) an **lcd_screen** (**12**), **rubber** (**20**) buttons, colorful **rubber** (**20**) changeable skins, a kinetic **battery** (**18**), and possibly speech **recognition** (**8**) **if it is still within the budget to include it. Several of the features that the group** (**3**) had wanted to integrate into the **design** (**22**) were either too costly or unavailable due to new limitations from the factory. The **group** (**3**) had to change many of the original **design** (**22**) elements to an alternative.

6.5.2. mmr/ilp

Is it possible that when we open our **fliptop (3) shell** (**6**) its a little compact **mirror** (**5**) and when you press a **button** (**36**) it then goes onto the **phone (9) display** (**7**) th– the **remote_control (36) display (7) thing** (**44**). Is it possible just **as** (**2**) an **option** (**8**) when we open it up **people** (**20**) can use their fingers to press the **button** (**36**) or we have **inside** (**3**) like a small **pointer (3) thing** (**44**) when **people** (**20**) want to. So should we be thinking of using **something** (**22**) like that in our **remote_control (36) design** (**22**) too. Which was the major **thing** (**44**) that **people** (**20**) wanted **market_research** (**15**). Not the actual **plastic** (**8**) outside **case** (**11**) just the **rubber (20) thing** (**44**) that goes round the outside. Some **kind** (**11**) of **thing** (**44**) or it gives a b – bleep **sound** (**2**) or some **kind** (**11**) of **sound** (**2**). *So f– on the s– simpler* **board** (**9**) *on the* **top** (**4**) *we have this* **button (36) rubber** (**20**) *buttons to keep frequently changing the channels.* It is not a **thing** (**44**) that **people** (**20**) are looking for. *We decided on the most* **important_aspect** (**6**) *I required in a* **remote_control (36) device** (**4**). And **rubber (20) as** (**2**) a padding or for the **grip (2) something** (**22**) like to add to the **design** (**22**). Well it is a **remote_control** (**36**). They also want a **remote_control (36) to be technologically innovative. First thing** (**44**) is basically on **design** (**22**). It is not **something** (**22**) that is come up in any of our focus groups and **market_research** (**15**). I do not see why the curved **thing** (**44**) is a **problem** (**4**). And second **thing** (**44**) is there is too much of confusion here. Ye– yeah I think I th– g– y– you could have a dual power **thing** (**44). So I think that is quite a flexible thing** (**44**). Icons or **something** (**22**) y– you have is a good **example** (**5**) of gui **graphic_user_interface** (**8**). *And second* **thing** (**44**) *is cer– certain standard buttons we should have.*

6.5.3. concept/ilp

The minutes from the last **time** (**11**). So we decided on our **market** (**12**). And so this **feedback** (**5**) from the **marketing_department** (**6**) is really about **trend** (**5**) watching. I'm w– I'm sorry. *We decided on the most* **important_aspect** (**6**) *I required in a* **remote_control (36) device** (**4**). Now the **fashion_update** (**6**) which relates to very personal preferences among our **subject_group** (**6**). And then we we're loo– looking into **battery** (**18**) options. I saw the standard **double_a** (**9**) and **triple_a** (**12**). And **dynamo** (**3**) might take **more_space** (**6**). It is moving a **lot** (**4**) of the **time** (**11**). It is 12 **point** (**6**) f–. Because we do not want customers to be like you know **charging** (**4**) like a **mobile_phone** (**18**) every **day** (**4**). If you had **something** (**22**) du– using the standard batteries and the **solar_charging** (**9**). The eternal battle for **control** (**19**) of the controls. Most current remotes use this silicone **pcb (2) board** (**9**) which pr– printed **circuit_board** (**15**). So is that **feature** (**10**) available in like **titanium** ((**7**). I know we were planning to do some **sort** (**10**) of **touch ((6) screen** (**9**). And g– **graphic_user_interface** (**8**). *So f– on the s– simpler* **board** (**9**) *on the* **top** (**4**) *we have this* **button (36) rubber** (**20**) *buttons to keep frequently changing the channels.* Is not that the **idea** (**8**). **Example** (**5**) the **volume** (**6**) and **channel (8) control** (**19**) buttons. Okay we had a latest **finding** (**6**) of **voice_recognition** (**21**). *And second* **thing** *(***44***) is cer– certain standard buttons we should have.* The **lcd** (**11**)'s not cheap. For the **body (6) design** (**22**) I think **plastic** (**8**). If we have got a **kind** (**11**) of **different_shape** (**12**) anyway. Which was the major **thing** (**44**) that **people** (**20**) wanted **market_research** (**15**). We are gonna use **fruit** (**7**) and **vegetable** (**4**) colours for the **rubber (20) cover** (**3**) the **case** (**11**) itself is **plastic** (**8**). So are we looking at **voice** (**8). But it is a good_idea** (**9**). I know at the **last_meeting** (**12**) we spoke about a **beeper** (**2**).

7. Conclusion and outlook

In this article, we provided an extensive comparison of global sentence and concept-based models for meeting summarization. The former give relevance and redundancy scores to each sentence selected for a summary while the later assess the relevance of sub-sentence units (called concepts) contained in a summary without explicitly modeling redundancy. In our experiments, concept-based models yield best results both in term of summary quality and in term of run time.

Though (greedy) sentence-based models were successfully used in the past, it seems that their global formulations do not provide the expected performance gain, and present excessive computation complexity. The use of ILP for optimizing global criteria is relatively new in the summarization community, and not all performance issues are fully understood. However, the run times of the sentence-based models addressed in this work can be explained by the high number of utterances showing same or similar relevance and redundancy, leading to possible solutions with the same objective function value that are exhaustively enumerated by the solver. In addition, the chosen similarity measures might not be the most appropriate for global models even though they were proved to work well with MMR. The similarity measures also share the keyphrases and the underlying idea with the concept-based model, ensuring a fair comparison.

Beside better performance and scalability, the concept-based approach is not affected by long ILP runtimes and provides greedy performance significantly better than the sentence-based models. The concept-based model can also be used for interactive summarization where the user is allowed to refine the set of concepts and their weight so that they are more relevant to his needs.

Using ASR instead of manual transcripts results in a uniform loss of performance for all systems, none of which seems more affected than the others. The ASR summaries may contain misrecognized words which are then compared to the human abstracts using ROUGE. That is, even if the selection is perfect in case of manual transcripts, the ROUGE score would be lower as it is based on exact word overlap. If the system were used by a human, this problem can be avoided by presenting the extracts in form of audio. Even though the quality of summaries will be improved by better speech recognition, the use of ASR confidence scores might help summarization systems when difficult acoustic conditions occur.

The quality of the chosen concepts is crucial – for both models. They need to be on the one hand informative and on the other hand representatively weighted according to their importance. A possible drawback of the current concept-based formulation is that each concept is only accounted for once. Experiments in (Gillick et al., 2009) revealed that for meeting summarization, it might be of interest to explore different ways of accounting for concepts. For example, allowing multiple occurrences per concept (e.g., once per speaker) as it might be the topic of a controversial discussion, thus all utterances containing it are of interest. A related question is, whether or not utterances with a semantic dependency to another (such as question answer pairs) should always be extracted as a combined unit. Though this sounds very reasonable, it is hard to realize for a general, broad summary (of fixed length) where one seeks to include as many topics as possible instead of lesser but more informative parts.

Although it is not the focus of this work, the presented keyphrase algorithm can be greatly enhanced using external world knowledge. For example, the meeting agenda (if available), information about the attendants and notes brought to or acquired during the meeting can be used to identify and weight concepts or complete utterances.

Surely, the concept-based formulation is not exploited to its full extent. Beside extending the concept idea as mentioned above, one could think of integrating some utterance level scores (e.g., grammaticality, automatic speech recognition confidence, length or number of concepts contained) directly to the optimization problem. This can help avoiding the inclusion of short, ill-formed or aborted utterances containing high value keyphrases. Xie et al. (2009a) introduced a first step towards augmenting the concept-based algorithm by integrating sentence weights. Gillick and Favre (2009) extended the original formulation to incorporate possible sentence compression. A promising summarization method proposed in (Lin et al., in press) shows that greedy solutions in summarization can lead to quasi-optimality when the objective function is submodular. It will be very interesting to merge the speed of that approach with the expressiveness of the ILP to combines the strengths of both approaches in one optimization.

As the use of acoustic and prosodic information helps with almost all speech-related tasks, it should also be integrated into the concept based system. A straight-forward way is to modify the concept/keyphrase weight according to information like fluency, sentence accent or utterance type (e.g., question vs. answer). Another, more flexible way is to attribute certain concepts to sentences based on acoustic or prosodic information, such as a disfluency score, utterance type. At a higher level, information describing how confident a speaker was could add to the reliability or trustworthiness of keyphrases. The probably most interesting aspect of using acoustic information is speech summarization without ASR by identifying frequent acoustic patterns, as for example in (Zhu et al., 2009), and use them as concepts.

References

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining Knowl. Discovery 2 (2), 121–167.

Carbonell, J., Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 335–336.

Christensen, H., Kolluru, B., Gotoh, Y., Renals, S., 2004. From text summarisation to style-specific summarisation for broadcast news. Lect. Notes Comput. Sci. 2997, 223–237.

Filatova, E., Hatzivassiloglou, V., 2004. Event-based extractive summarization. In: Proc. ACL Workshop on Summarization.

Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C., 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Trans. Speech Audio Process. 12 (4), 401–408.

Garg, N., Riedhammer, B.F.K., Hakkani-Tür, D., 2009. ClusterRank: a graph based method for meeting summarization. In: Proc. Annual Conf. of the Internat. Speech Communication Association (INTERSPEECH), pp. 1499–1502.

Gillick, D., Favre, B., 2009. A scalable global model for summarization. In: Proc. ACL-HLT Workshop on Integer Linear Programming for Natural Language Processing, pp. 10–18.

Gillick, D., Favre, B., Hakkani-Tür, D., 2008. The ICSI Summarization System at TAC'08. In: Proc. of the Text Analysis Conf. Workshop, pp. 227–234.

Gillick, D., Riedhammer, K., Favre, B., Hakkani-Tür, D., 2009. A global optimization framework for meeting summarization. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4769–4772.

Ha, L., Sicilia-Garcia, E., Ming, J., Smith, F., 2002. Extension of Zipf's law to words and phrases. In: Proc. Internat. Conf. on Computational Linguistics, pp. 1–6.

Hori, C., Furui, S., 2000. Improvements in automatic speech summarization and evaluation methods. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), pp. 326–329.

Hori, C., Furui, S., Malkin, R., Yu, H., Waibel, A., 2002. Automatic speech summarization applied to English broadcast news speech. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 9–12.

Hovy, E., Lin, C., Zhou, L., Fukumoto, J., 2006. Automated summarization evaluation with basic elements. In: Proc. Internat. Conf. on Language Resources and Evaluation (LREC).

Huang, Z., Harper, M., Wang, W., 2007. Mandarin part-of-speech tagging and discriminative reranking. In: Proc. EMNLP/CoNLL, pp. 1093–1102.

Inoue, A., Mikami, T., Yamashita, Y., 2004. Improvement of speech summarization using prosodic information. In: Proc. Internat. Conf. on Speech Prosody, pp. 599–602.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al., 2003. The ICSI meeting corpus. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 364–367.

Lin, C., 2004. ROUGE: a package for automatic evaluation of summaries. In: Proc. Workshop on Text Summarization Branches Out (WAS), pp. 25–26.

Lin, H., Bilmes, J., Xie, S., in press. Graph-based submodular selection for extractive summarization. In: Proc. IEEE Workshop on Speech Recognition and Understanding (ASRU). Q1

Liu, F., Liu, Y., 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In: Proc. ACL-HLT, pp. 201–204.

Liu, F., Liu, Y., 2009. From extractive to abstractive meeting summaries: can it be done by sentence compression? In: Proc. ACL-IJCNLP (short paper), pp. 261–264.

Liu, Y., Xie, S., 2008. Impact of automatic sentence segmentation on meeting summarization. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5009–5012.

Liu, F., Liu, F., Liu, Y., 2008. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In: Proc. IEEE Workshop on Spoken Language Technologies (SLT), pp. 181–184.

Liu, F., Pennell, D., Liu, F., Liu, Y., 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In: Proc. ACL-HLT, pp. 620–628.

Maskey, S., Hirschberg, J., 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In: Proc. European Conf. on Speech Communication and Technology (EUROSPEECH), pp. 621–624.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al., 2005. The AMI meeting corpus. In: Proc. of Measuring Behavior.

McDonald, R., 2007. A study of global inference algorithms in multi-document summarization. Lect. Notes Comput. Sci. 4425, 557–564.

Mieskes, M., Mller, C., Strube, M., 2007. Improving extractive dialogue summarization by utilizing human feedback. In: Proc. Artificial Intelligence and Applications (AIA), pp. 627–632.

Mrozinski, J., Whittaker, E., Chatain, P., Furui, S., 2005. Automatic sentence segmentation of speech for automatic summarization. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 981–984.

Please cite this article in press as: Riedhammer, K., et al. Long story short – Global unsupervised models for keyphrase based meeting summarization. Speech Comm. (2010), doi:10.1016/j.specom.2010.06.002

Murray, G., Renals, S., 2007. Term-weighting for summarization of multi-party spoken dialogues. In: Proc. ACM Workshop on Machine Learning for Multimodal Interaction, pp. 156–167.

Murray, G., Renals, S., Carletta, J., 2005a. Extractive summarization of meeting recordings. In: Proc. European Conf. on Speech Communication and Technology (EUROSPEECH), pp. 593–596.

Murray, G., Renals, S., Carletta, J., Moore, J., 2005b. Evaluating automatic summaries of meeting recordings. In: Proc. ACL Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 33–40.

Murray, G., Renals, S., Carletta, J., Moore, J., 2006. Incorporating speaker and discourse features into speech summarization. In: Proc. ACL-HLT, pp. 367–374.

Murray, G., Kleinbauer, T., Poller, P., Renals, S., Kilgour, J., Becker, T., 2008. Extrinsic summarization evaluation: a decision audit task. In: Proc. Internat. Workshop on Machine Learning for Multimodal Interaction (MLMI), pp. 349–360.

Nenkova, A., Passonneau, R., 2004. Evaluating content selection in summarization: the pyramid method. In: Proc. Joint Annual Meeting of HLT/NAACL.

Penn, G., Zhu, X., 2008. A critical reassessment of evaluation baselines for speech summarization. In: Proc. ACL-HLT, pp. 470–478.

Renals, S., Hain, T., Bourlard, H., 2007. Recognition and interpretation of meetings: the AMI and AMIDA projects. In: Proc. IEEE Workshop on Speech Recognition and Understanding (ASRU).

Riedhammer, K., Favre, B., Hakkani-Tür, D., 2008a. A keyphrase based approach to interactive meeting summarization. In: Proc. IEEE Workshop on Spoken Language Technologies (SLT), pp. 153–156.

Riedhammer, K., Gillick, D., Favre, B., Hakkani-Tür, D., 2008b. Packing the meeting summarization knapsack. In: Proc. Annual Conf. of the Internat. Speech Communication Association (INTERSPEECH), pp. 2434–2437.

Santorini, B., 1990. Part-of-Speech Tagging Guidelines for the Penn Treebank Project (third revision). Tech. Rep. MS-CIS-90-47, University of Pennsylvania, Department of Computer and Information Science.

Schapire, R.E., Singer, Y., 2000. BoosTexter: a boosting-based system for text categorization. Machine Learn. (39), 135–168.

Takamura, H., Okumura, M., 2009. Text summarization model based on maximum coverage problem and its variant. In: Proc. Conf. of the European Chapter of the ACL, pp. 781–789.

Thede, S., Harper, M., 1999. A second-order hidden Markov model for part-of-speech tagging. In: Proc. ACL, pp. 175–182.

Xie, S., Favre, B., Hakkani-Tür, D., Liu, Y., 2009a. Leveraging sentence weights in a concept-based optimization framework for meeting summarization. In: Proc. Annual Conf. of the Internat. Speech Communication Association (INTERSPEECH), pp. 1503–1506.

Xie, S., Hakkani-Tür, D., Favre, B., Liu, Y., 2009b. Integrating prosodic features in extractive meeting summarization. In: Proc. IEEE Workshop on Speech Recognition and Understanding (ASRU).

Zechner, K., 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. Comput. Linguist. 28 (4), 447–485.

Zhang, J., Fung, P., 2007. Speech summarization without lexical features for Mandarin broadcast news. In: Proc. ACL-HLTH (short Paper), pp. 213–216.

Zhu, X., Penn, G., 2006. Utterance-level extractive summarization of open-domain spontaneous conversations with rich features. In: IEEE Internat. Conf. on Multimedia and Expo, pp. 793–796.

Zhu, Q., Stolcke, A., Chen, B., Morgan, N., 2005. Using MLP features in SRI's conversational speech recognition system. In: Proc. European Conf. on Speech Communication and Technology (EUROSPEECH), pp. 2141–2144.

Zhu, X., Penn, G., Rudzicz, F., 2009. Summarizing multiple spoken documents: Finding evidence from untranscribed audio. In: Proc. Internat. Joint Conference on Natural Language Processing of the AFNLP, pp. 549–557.