Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization

Guokan Shang[1,2], Wensi Ding[1], Zekun Zhang[1], Antoine J.-P. Tixier[1], Polykarpos Meladianos[1,3], Michalis Vazirgiannis[1,3], Jean-Pierre Lorré[2]

[1]École Polytechnique, [2]Linagora, [3]AUEB

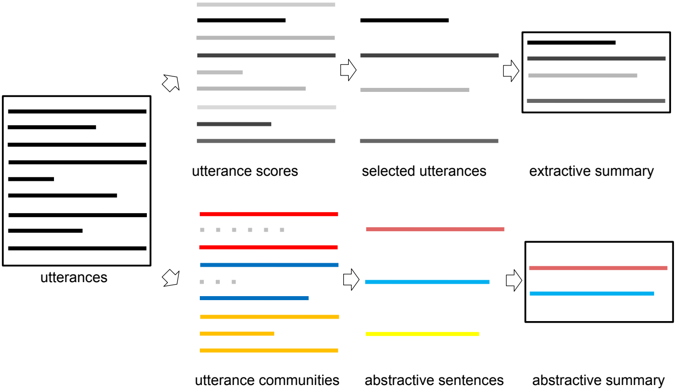
LaTeXof the slides: `https://www.overleaf.com/read/dscfmyynczjv`

Table of Contents

Introduction

Spontaneous multi-party meeting speech transcription is made of often ill-formed and ungrammatical text fragments (called *utterances*).

⇒ *Summarizing transcription requires approaches that differ from traditional document summarization.*



⇒ *Abstractive summaries are preferred to extractive ones by human judges. (Murray, Carenini, and Ng 2010)*

Related work

Our system builds on 4 pieces of work:

- Filippova 2010 *multi-sentence compression*
  - unsupervised, simple approach based on word graph
  - edge-weights → k-shortest paths → heuristics and re-ranking
- Boudin and Morin 2013 *keyphrase extraction*
  - same as Filippova 2010 +
  - re-ranking taking into account information coverage (TextRank scores)
  - account for punctuation

- Mehdad et al. 2013 *abstractive meeting summarization*
  - community detection
  - each community is fused with Filippova's approach +
  - WordNet to capture synonymy and hypo/hypernymy when building graph
  - re-ranking taking into account information coverage (TF-IDF scores) and grammaticality (via a language model)

- Tixier et al. 2017 *extractive meeting summarization*
  - submodularity
  - coverage term based on k-core decomposition of graph-of-words

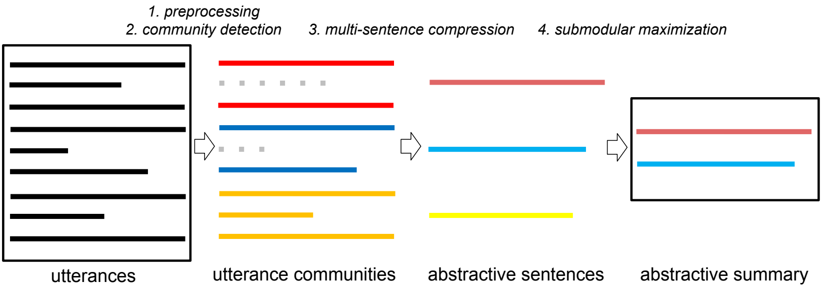
Contributions

Our main contributions

Capitalizing on the respective strengths of the 4 aforementioned approaches while addressing their weaknesses:

1. fully unsupervised meeting summarization framework instead of the supervised one proposed in Mehdad et al. 2013
2. novel edge weight assignment based on word embeddings and path re-ranking strategy (fluency, coverage and diversity) for word graph
   - fluency with a LM like in Mehdad et al. 2013
   - coverage and punctuation like in Boudin and Morin 2013 but better coverage based on graphs-of-words and degeneracy
   - diversity based on word embeddings
3. final summary constructed with submodularity

Pipeline


1. preprocessing
2. community detection
3. multi-sentence compression
4. submodular maximization
utterances
utterance communities
abstractive sentences
abstractive summary

1. Text Preprocessing & 2. Utterance Community Detection

Preprocessing

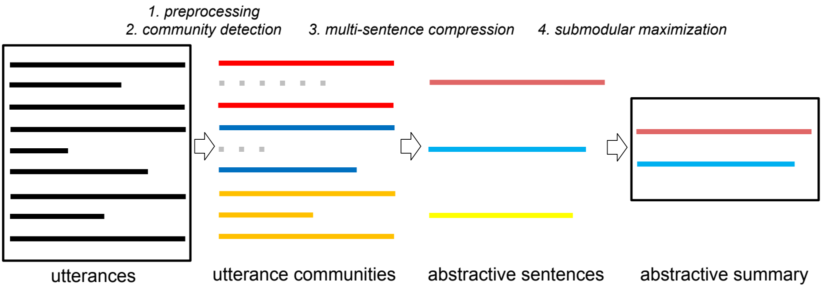
- Initial ellipsis: *'kay, 'til, 'em → okay, until, them*
- Consecutive repeated unigram and bigram terms: *remote control remote control → remote control*
- ASR tags are filtered out: *<vocalsound>*
- Filler words are discarded: *uh-huh, okay well, by the way*
- Consecutive stopwords at head and tail of utterance are stripped
- Utterances containing less than 3 non-stopwords are pruned out

Clustering

Group together the utterances that should be summarized by a common abstractive sentence. (Murray, Carenini, and Ng 2012)

1. Utterances → TF-IDF weight matrix
2. Latent Semantic Analysis
3. K-means algorithm (on the SVD result) - 35 to 50 clusters

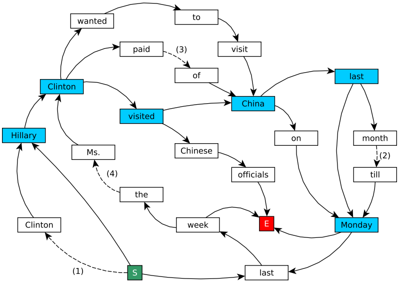
Pipeline


1. preprocessing
2. community detection
3. multi-sentence compression
4. submodular maximization
utterances
utterance communities
abstractive sentences
abstractive summary

3. Multi-Sentence Compression Graph (MSCG)

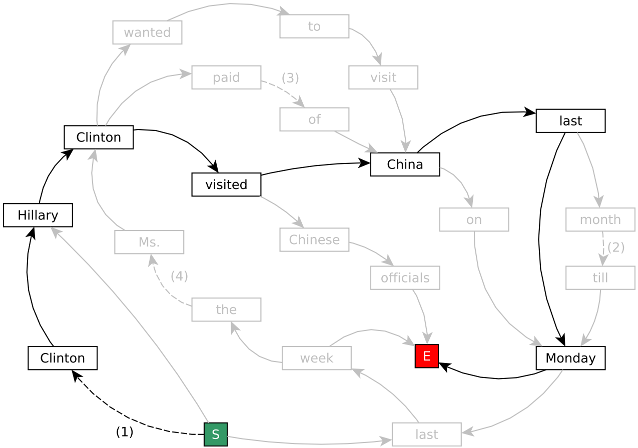
Generate an abstractive sentence for each utterance community with MSCG.

1. The wife of a former U.S. president Bill Clinton Hillary Clinton visited China last Monday
2. Hillary Clinton wanted to visit China last month but postponed her plans till Monday last week
3. Hillary Clinton paid a visit to the People Republic of China on Monday
4. Last week the Secretary of State Ms. Clinton visited Chinese officials



⇒ *Redundancy provides a reliable way of generating grammatical sentences.* (Filippova 2010)

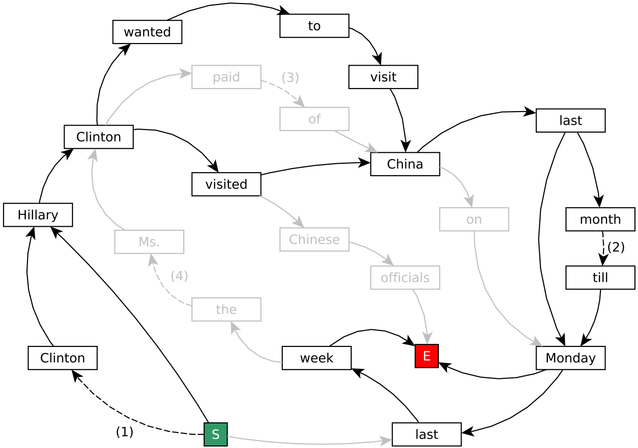
3.1. MSCG Building (1/4)



(1) *The wife of a former U.S. president Bill* Clinton Hillary Clinton visited China last Monday[1]

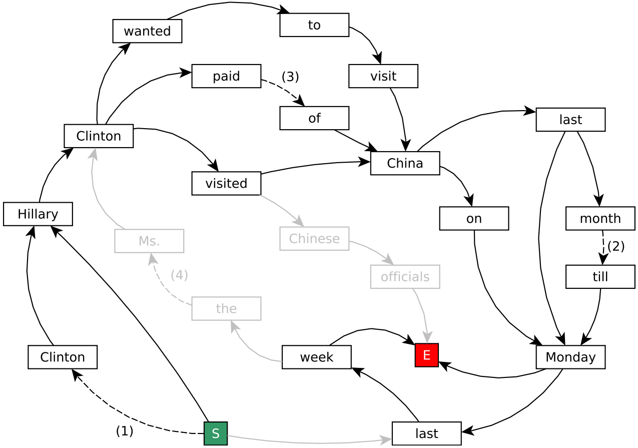
[1] Italicized fragments from the sentences are replaced with dashed arrow for clarity in the graph.

3.1. MSCG Building (2/4)



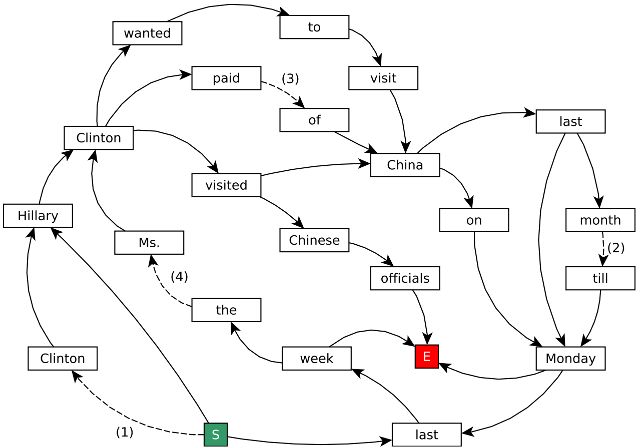
(2) Hillary Clinton wanted to visit China last month *but postponed her plans* till Monday last week

3.1. MSCG Building (3/4)



(3) Hillary Clinton paid *a visit to the People Republic of* China on Monday

3.1. MSCG Building (4/4)

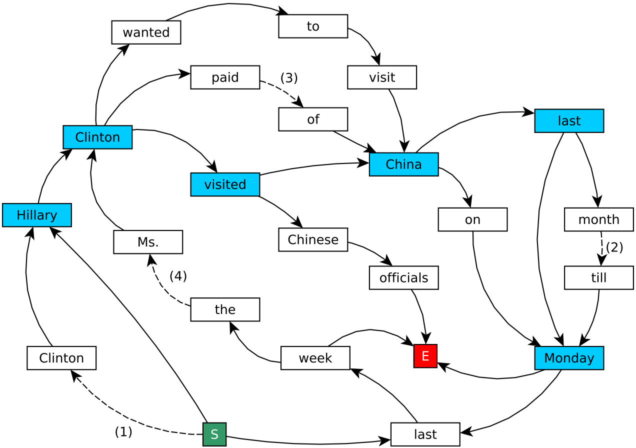


(4) Last week the *Secretary of State* Ms. Clinton visited Chinese officials

⇒ *Every input sentence corresponds to a loopless path in the graph.*

⇒ *There are many other paths.*

3.1. Objective of MSCG Building



⇒ *Find the best compression path:* ***Hilary Clinton visited China last Monday***.

3.2. Edge Weight Assignment

- **Local co-occurrence statistics** (Filippova 2010):

$$w'(p_i, p_j) = \frac{freq(p_i) + freq(p_j)}{\sum_{P \in G', p_i, p_j \in P} \mathrm{diff}(P, p_i, p_j)^{-1}} \quad (1)$$

Favors edges between words that frequently appear close to each other (*word association*).

$freq(p_i)$: number of words mapped to the node p_i.
$diff(P, p_i, p_j)^{-1}$: inverse of the distance between p_i and p_j in path P.

- **Global exterior knowledge**: Word Attraction Score (Wang, Liu, and McDonald 2014):

$$w''(p_i, p_j) = \frac{freq(p_i) \times freq(p_j)}{d^2_{p_i, p_j}} \quad (2)$$

Favor paths going through salient nodes that are close in the embedding space (*semantic relatedness*).

d_{p_i, p_j}: Euclidean distance of the word embedding vectors for p_i and p_j.

- **Final edge weight** (the lower the better):

$$w'''(p_i, p_j) = \frac{w'(p_i, p_j)}{w''(p_i, p_j)} \quad (3)$$

3.3. Path Selection and Reranking (1/2)

- Path score as its **cumulative edge weights** (the lowest is the best compression path):

$$W(P) = \sum_{i=1}^{|P|-1} w'''(p_i, p_{i+1}) \tag{4}$$

Reranking

The path with the lowest score does not guarantee its readability nor informativeness. (Boudin and Morin 2013)

$\Rightarrow$ *Reranking N best paths is necessary.*

3.3. Path Selection and Reranking (2/2)

- **Fluency** (Mehdad et al. 2013): estimate readability of MSCG path P based on a 3-gram language model

$$F(P) = \frac{\sum_{i=1}^{|P|} \log Pr(p_i | p_{i-n+1}^{i-1})}{\#n\text{-}gram} \quad (5)$$

- **Coverage** (Mehdad et al. 2013): estimate the information covered by P

$$C(P) = \frac{\sum_{p_i \in P} \text{TW-IDF}(p_i)}{\# p_i} \quad (6)$$

  TW: term CoreRank score of p_i in the GoW of the community. (Tixier, Malliaros, and Vazirgiannis 2016)

- **Diversity**: estimate the diversity of the information contained by P

$$D(P) = \frac{\sum_{j=1}^{k} 1_{\exists p_i \in P | p_i \in \text{cluster}_j}}{|P|} \quad (7)$$

  The number of different word clusters covered by the path

- **Final path score**: select the path with the lowest score per community

$$\text{score}(P) = \frac{W(P)}{|P| \times F(P) \times C(P) \times D(P)} \quad (8)$$

Diversity

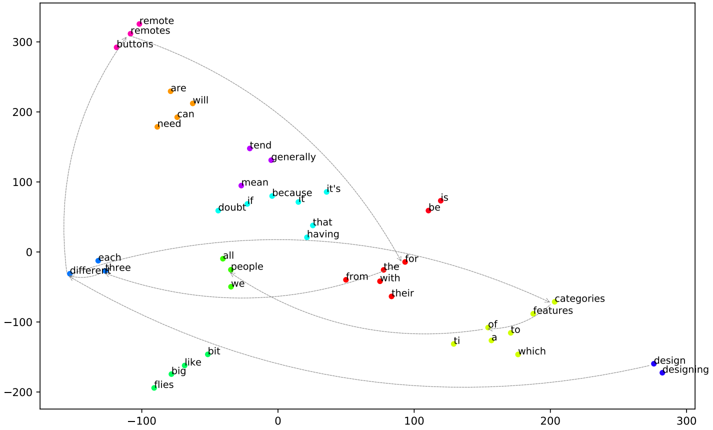
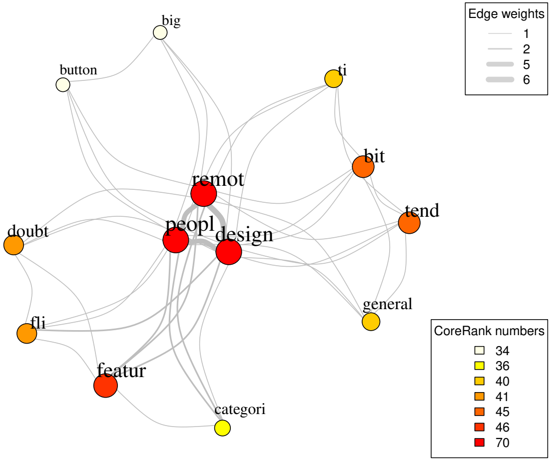
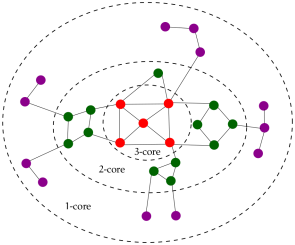


Figure: t-SNE visualization of the GoogleNews vectors of the words in an utterance community. Arrows join the words in the best compression path. Movements in the embedding space, as measured by the number of unique clusters covered by the path (here, 6/11), can provide a sense of the diversity of the compressed sentence, as formalized in Equation 7.

Background on Keyword Extraction with Graph-of-words and CoreRank

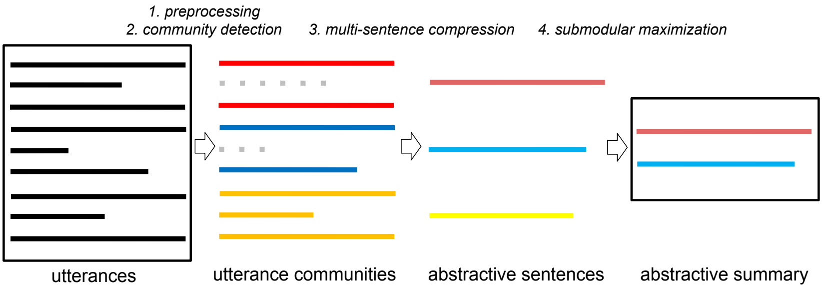
$$TW\text{-}IDF(t,d,D) = TW(t,d) \times IDF(t,D) \tag{9}$$





(Rousseau and Vazirgiannis 2015; Tixier, Malliaros, and Vazirgiannis 2016; Meladianos et al. 2017)

Pipeline


1. preprocessing
2. community detection
3. multi-sentence compression
4. submodular maximization
utterances
utterance communities
abstractive sentences
abstractive summary

4. Budgeted Submodular Maximization

Generate the final summary by selecting an optimal subset S from the set of abstractive sentences $\mathcal{S}$ under a budget constraint.

$$\boxed{\underset{S \subseteq \mathcal{S}}{\operatorname{argmax}}\, f(S) | \sum_{s \in S} cost_s \leq Budget}$$

NP-hard, but near-optimal performance can be guaranteed with a modified greedy algorithm (H. Lin and Bilmes 2010) that iteratively selects the sentence s that maximizes the ratio of summary quality function gain to scaled cost ${}^{f(G \cup s) - f(G)}/_{cost_s^r}$ (where G is the current subset and $r \geq 0$ is a scaling factor).

Summary quality function f is non-decreasing and **submodular** taking both coverage and diversity into account:

$$f(S) = c(S) + \lambda d(S)$$

$$c(S) = \sum_{s_i \in S} n_{s_i} w_{s_i}, \; d(S) = \sum_{j=1}^{k} 1_{\exists s_i \in S, s_i \in cluster_j}$$

$\lambda \geq 0$: trade-off parameter, n_{s_i}: number of occurrences of word s_i in S, w_{s_i}: CoreRank score of word s_i

Background on Submodularity

- **Submodularity** (Krause and Golovin 2014):

  A set function $F : 2^V \to \mathcal{R}$ where $V = \{v_1, ..., v_n\}$ is said to be *submodular* if it satisfies the property of *diminishing returns*:

$$\forall A \subseteq B \subseteq V \backslash v,$$
$$F(A \cup v) - F(A) \geq F(B \cup v) - F(B)$$

  the gain of adding a new sentence to a given summary should be greater than the gain of adding the same sentence to a larger summary containing the smaller one

  the set function $F(\cdot)$ is *monotone non-decreasing*:

$$\forall A \subseteq B, F(A) \leq F(B)$$

  the quality of a summary can only increase or stay the same as it grows in size, i.e., as we add sentences to it

Reference systems

Baselines

- **Random** random selection of utterances until budget is violated (30 runs)
- **Longest Greedy** longest utterance selected at each step until budget is violated
- **TextRank** Mihalcea and Tarau 2004
- **ClusterRank** Garg et al. 2009
- **Oracle** Tixier et al. 2017
- **CoreRank Submodular & PageRank Submodular** Tixier et al. 2017

Variants of our system

- in word graph: simple edge weights with basic re-ranking (based on path length), like in **Filippova 2010**
- Filippova's edge weights + coverage scores of keyphrases (based on TextRank), like in **Boudin and Morin 2013**
- Filippova's edge weights + with re-ranking taking into account length and coverage scores of keyphrases (based on TextRank), like in **Boudin and Morin 2013**
- $\sim$ Filippova's edge weights + length, fluency and coverage scores (based on TFIDF scores of nouns), like in **Mehdad et al. 2013**

Datasets & Metrics

Datasets

- **AMI Corpus**[2]
  - Role-play meetings of participants within a fictive company
  - 47 for development, 20 for test
  - Each meeting is associated with one human-written abstractive summary
- **ICSI Corpus**[3]
  - Real life meetings
  - 25 for development, 6 for test
  - Each meeting is associated with three human-written abstractive summaries

Metrics

ROUGE-1, ROUGE-2 and ROUGE- SU4 metrics (C.-Y. Lin 2004), respectively based on unigram, bigram, and unigram plus skip-bigram overlap with maximum skip distance of 4.

[2] `http://groups.inf.ed.ac.uk/ami/corpus/index.shtml`
[3] `http://groups.inf.ed.ac.uk/ami/icsi/index.shtml`

Parameter tuning

Grid Search		
step 2	**#communities** n	[20, 60] with step size 5
step 3	**minimum path length (#words)** z	[6, 16] with step size 2
step 4	**lambda** λ	[0, 1] with step size 0.1
	scaling factor r	[0, 2] with step size 0.1

$\Rightarrow$ We optimize parameters over summarization size: 350 words for AMI, 450 words for ICSI.

System	AMI	ICSI
Our System	50, 8, (0.7, 0.5)	40, 14, (0.0, 0.0)
Our System (Baseline)	50, 12, (0.3, 0.5)	45, 14, (0.1, 0.0)
Our System (KeyRank)	50, 10, (0.2, 0.9)	45, 12, (0.3, 0.4)
Our System (FluCovRank)	35, 6, (0.4, 1.0)	50, 10, (0.2, 0.3)

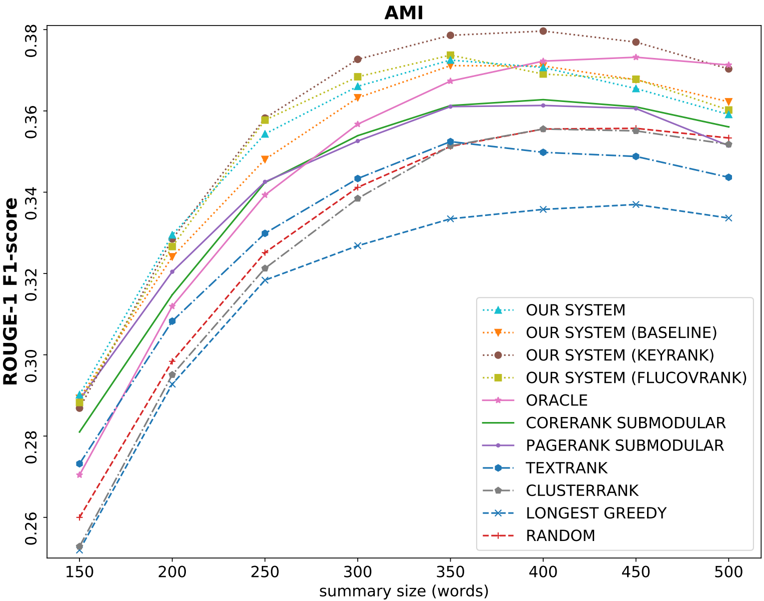
Table: Optimal parameter values $n, z, (\lambda, r)$.

ROUGE Results AMI

	AMI ROUGE-1			AMI ROUGE-2			AMI ROUGE-SU4		
	R	P	F-1	R	P	F-1	R	P	F-1
Our System	41.83	34.44	37.25	8.22	6.95	7.43	15.83	13.70	14.51
Our System (Baseline)	41.56	34.37	37.11	7.88	6.66	7.11	15.36	13.20	14.02
Our System (KeyRank)	42.43	35.01	**37.86**	8.72	7.29	**7.84**	16.19	13.76	**14.71**
Our System (FluCovRank)	41.84	34.61	37.37	8.29	6.92	7.45	16.28	13.48	14.58
Oracle	40.49	34.65	**36.73**	8.07	7.35	**7.55**	15.00	14.03	**14.26**
CoreRank Submodular	41.14	32.93	36.13	8.06	6.88	7.33	14.84	13.91	14.18
PageRank Submodular	40.84	33.08	36.10	8.27	6.88	7.42	15.37	13.71	14.32
TextRank	39.55	32.60	35.25	7.67	6.43	6.90	14.87	12.87	13.62
ClusterRank	39.36	32.53	35.14	7.14	6.05	6.46	14.34	12.80	13.35
Longest Greedy	37.31	30.93	33.35	5.77	4.71	5.11	13.79	11.11	12.15
Random	39.42	32.48	35.13	6.88	5.89	6.26	14.07	12.70	13.17

Table: Macro-averaged results for 350 word summaries (ASR transcriptions).

ROUGE-1 F1-score

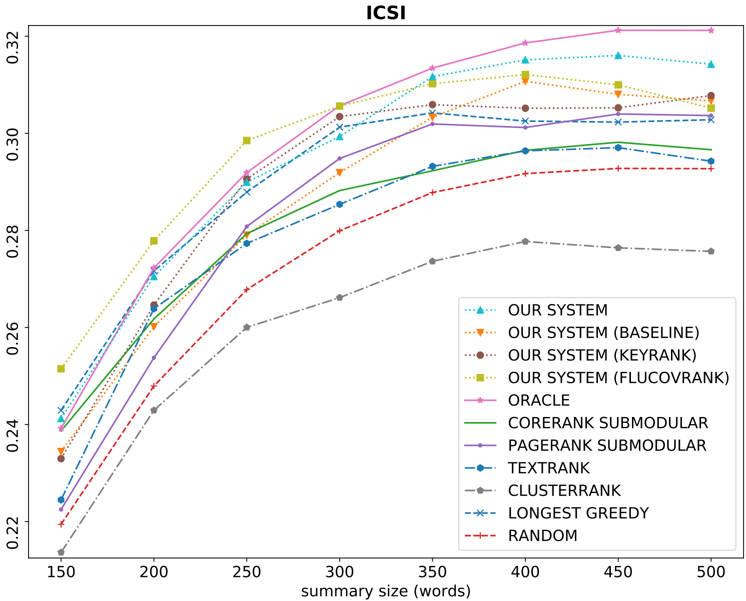


ROUGE Results ICSI

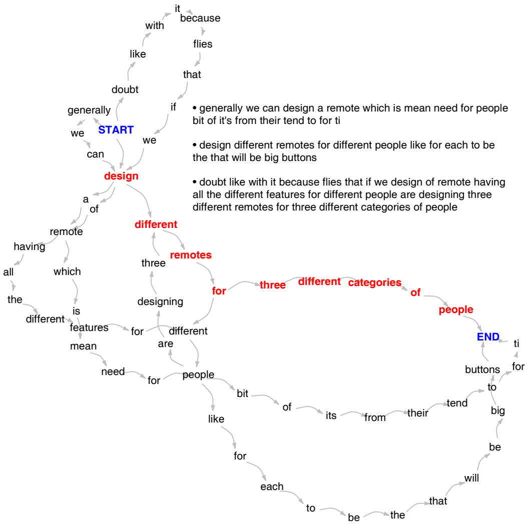
	ICSI ROUGE-1			ICSI ROUGE-2			ICSI ROUGE-SU4		
	R	P	F-1	R	P	F-1	R	P	F-1
Our System	36.99	28.12	**31.60**	5.41	4.39	4.79	13.10	10.17	**11.35**
Our System (Baseline)	36.39	27.20	30.80	5.19	4.12	4.55	12.59	9.70	10.86
Our System (KeyRank)	35.95	27.00	30.52	4.64	3.64	4.04	12.43	9.23	10.50
Our System (FluCovRank)	36.27	27.56	31.00	5.56	4.35	**4.83**	13.47	9.85	11.29
Oracle	37.91	28.39	**32.12**	5.73	4.82	**5.18**	13.35	10.73	**11.80**
CoreRank Submodular	35.22	26.34	29.82	4.36	3.76	4.00	12.11	9.58	10.61
PageRank Submodular	36.05	26.69	30.40	4.82	4.16	4.42	12.19	10.39	11.14
TextRank	34.89	26.33	29.70	4.60	3.74	4.09	12.42	9.43	10.64
ClusterRank	32.63	24.44	27.64	4.03	3.44	3.68	11.04	8.88	9.77
Longest Greedy	35.57	26.74	30.23	4.84	3.88	4.27	13.09	9.46	10.90
Random	34.78	25.75	29.28	4.19	3.51	3.78	11.61	9.37	10.29

Table: Macro-averaged results for 450 word summaries (ASR transcriptions).

ROUGE-1 F1-score



Example



- generally we can design a remote which is mean need for people bit of it's from their tend to for ti
- design different remotes for different people like for each to be the that will be big buttons
- doubt like with it because flies that if we design of remote having all the different features for different people are designing three different remotes for three different categories of people

http://datascience.open-paas.org/abs_summ_app

Reference Summary AMI TS3003c

The project manager opened the meeting and recapped the decisions made in the previous meeting.
The marketing expert discussed his personal preferences for the design of the remote and presented the results of trend-watching reports, which indicated that there is a need for products which are fancy, innovative, easy to use, in dark colors, in recognizable shapes, and in a familiar material like wood.
The user interface designer discussed the option to include speech recognition and which functions to include on the remote.
The industrial designer discussed which options he preferred for the remote in terms of energy sources, casing, case supplements, buttons, and chips.
The team then discussed and made decisions regarding energy sources, speech recognition, LCD screens, chips, case materials and colors, case shape and orientation, and button orientation.
The team members will look at the corporate website.
The user interface designer will continue with what he has been working on.
The industrial designer and user interface designer will work together.
The remote will have a docking station.
The remote will use a conventional battery and a docking station which recharges the battery.
The remote will use an advanced chip.
The remote will have changeable case covers.
The case covers will be available in wood or plastic.
The case will be single curved.
Whether to use kinetic energy or a conventional battery with a docking station which recharges the remote.
Whether to implement an LCD screen on the remote.
Choosing between an LCD screen or speech recognition.

Example Summary AMI TS3003c manual transcription of Our System

attract elderly people can use the remote control
changing channels button on the right side that would certainly yield great options for the design of the remote
personally i dont think that older people like to shake your remote control
imagine that the remote control and the docking station
remote control have to lay in your hand and right hand users
finding an attractive way to control the remote control
casing the manufacturing department can deliver a flat casing single or double curved casing
top of that the lcd screen would help in making the remote control easier
increase the price for which were selling our remote control
remote controls are using a onoff button still on the top
apply remote control on which you can apply different case covers
button on your docking station which you can push and then it starts beeping
surveys have indicated that especially wood is the material for older people
mobile phones so like the nokia mobile phones when you can change the case
greyblack colour for people prefer dark colours
brings us to the discussion about our concepts
docking station and small screen would be our main points of interest
industrial designer and user interface designer are going to work
innovativeness was about half of half as important as the fancy design
efficient and cheaper to put it in the docking station
case supplement and the buttons it really depends on the designer
start by choosing a case
deployed some trendwatchers to milan

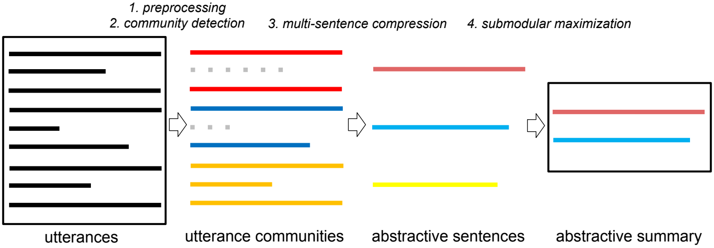
Next steps

Capitalize on Deep Learning

- to improve abstractive community detection step
- to improve language generation step
- to generate summaries in an end-to-end fashion

Leverage more data

- annotations and nonverbal information
- meeting metadata



References I

- Lin, Chin-Yew (2004). "Rouge: A package for automatic evaluation of summaries". In: **Text summarization branches out: Proceedings of the ACL-04 workshop**. Vol. 8. Barcelona, Spain.
- Mihalcea, Rada and Paul Tarau (2004). "Textrank: Bringing order into text". In: **Proceedings of the 2004 conference on empirical methods in natural language processing**.
- Garg, Nikhil et al. (2009). "Clusterrank: a graph based method for meeting summarization". In: **Tenth Annual Conference of the International Speech Communication Association**.
- Filippova, Katja (2010). "Multi-sentence compression: Finding shortest paths in word graphs". In: **Proceedings of the 23rd International Conference on Computational Linguistics**. Association for Computational Linguistics, pp. 322–330.
- Lin, Hui and Jeff Bilmes (2010). "Multi-document summarization via budgeted maximization of submodular functions". In: **Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics**. Association for Computational Linguistics, pp. 912–920.
- Murray, Gabriel, Giuseppe Carenini, and Raymond Ng (2010). "Generating and validating abstracts of meeting conversations: a user study". In: **Proceedings of the 6th International Natural Language Generation Conference**. Association for Computational Linguistics, pp. 105–113.

References II

- Murray, Gabriel, Giuseppe Carenini, and Raymond Ng (2012). "Using the omega index for evaluating abstractive community detection". In: **Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization**. Association for Computational Linguistics, pp. 10–18.
- Boudin, Florian and Emmanuel Morin (2013). "Keyphrase extraction for n-best reranking in multi-sentence compression". In: **North American Chapter of the Association for Computational Linguistics (NAACL)**.
- Mehdad, Yashar et al. (2013). "Abstractive Meeting Summarization with Entailment and Fusion.". In: **ENLG**, pp. 136–146.
- Krause, Andreas and Daniel Golovin (2014). **Submodular function maximization.**.
- Wang, Rui, Wei Liu, and Chris McDonald (2014). "Corpus-independent generic keyphrase extraction using word embedding vectors". In: **Software Engineering Research Conference**. Vol. 39.
- Rousseau, François and Michalis Vazirgiannis (2015). "Main core retention on graph-of-words for single-document keyword extraction". In: **European Conference on Information Retrieval**. Springer, pp. 382–393.
- Tixier, Antoine, Fragkiskos Malliaros, and Michalis Vazirgiannis (2016). "A graph degeneracy-based approach to keyword extraction". In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1860–1870.
- Meladianos, Polykarpos et al. (2017). "Real-Time Keyword Extraction from Conversations". In: **EACL 2017**, p. 462.

References III

- Shang, Guokan et al. (2018). "Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization". In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, pp. 664–674. URL: http://aclweb.org/anthology/P18-1062.