Research Institute for Linguistics

Hungarian Academy of Sciences

MŰEGYETEM 1782

Dept. of Telecommunications and Media Informatics

University of Technology and Economics

AUTOMATIC SUMMARIZATION OF HIGHLY SPONTANEOUS SPEECH

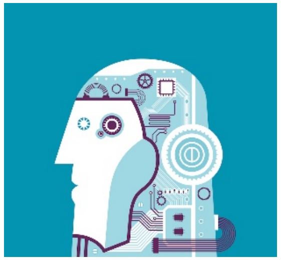
András Beke – György Szaszák

DATA EXPLOSION: HUMAN VS MACHINE PROCESSING

- The **UNSTUCTURED** data explosion
  - Growing 10x every 5 years and 100x every 10 years
  - Requires a new approach

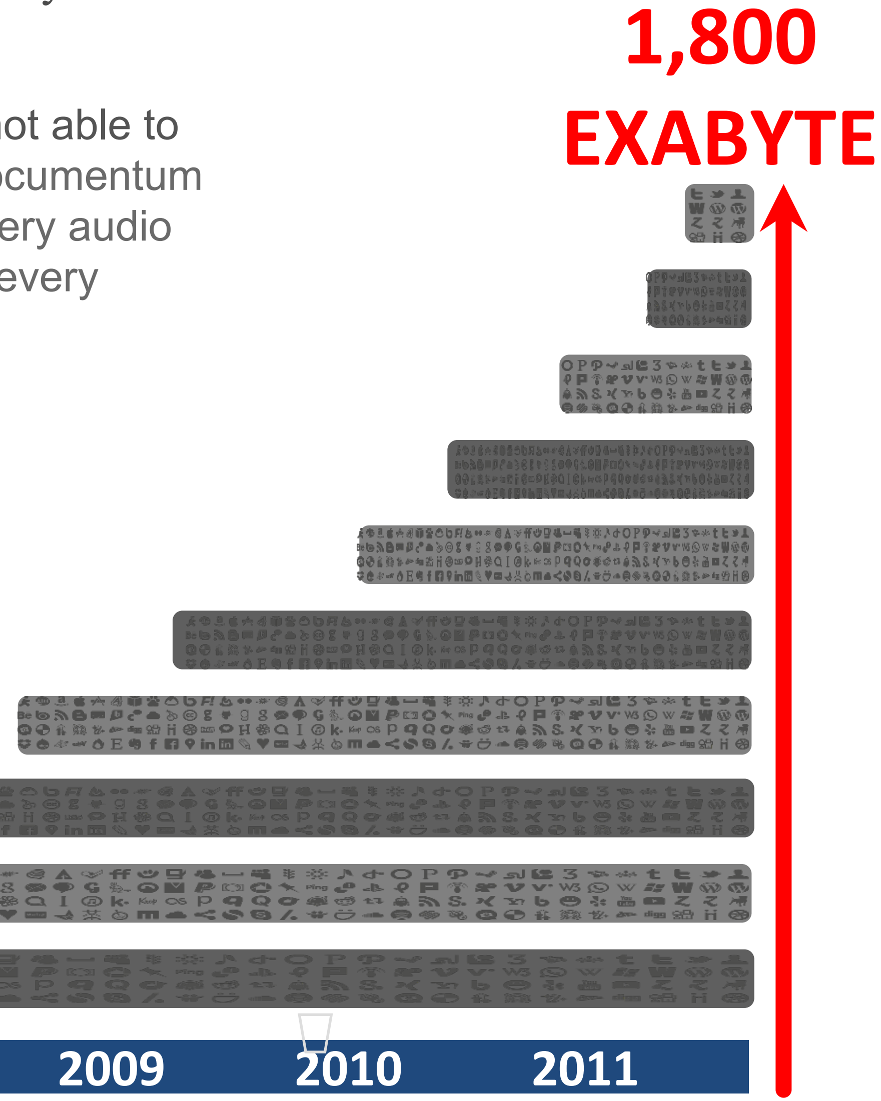
The **human** not able to read every documentum or to listen every audio file, or watch every movies!

STATISTICAL MACHINE LEARNING



STUCTURED data

180 EXABYTE



AUTOMATIC SUMMARIZATION

POSSIBLE APPROACHES

- information retrieval
- document clustering
- information extraction
- visualization
- question answering
- text summarization

TYPES OF SUMMARIZATION

- Indicative
  - Describes the document and its contents
- Informative
  - ‘Replaces’ the document
- Extractive
  - Concatenate pieces of existing document
- Generative
  - Creates a new document
- Document compression

COMPARING SPEECH AND TEXT SUMMARIZATION

ALIKE

- Identifying important information
- Some lexical, discourse features
- Extraction or generation or compression

DIFFERENT

– Speech Signal

– Prosodic features

– NLP tools?

– Segments – sentences?

– Generation?

– Errors

– Data size

Sentence Extraction/Similarity measures (Salton, et al. 1995)

- Extract sentences by their similarity to a topic sentence and their dissimilarity to sentences already in summary (Maximal Marginal Relativity)
- Similarity measures
  - Cosine Measure
  - Vocabulary Overlap
  - Topic word overlap
  - Content Signatures Overlap

- **Automatic sentence segmentation (tokenization) is crucial before such a sentence based extractive summarization (Liu and Xi 2008). The difficulty comes not only from recognition errors, but also from missing punctuation marks, which would be fundamental in syntactic parsing and POS tagging (disambiguation).**

Our work

- This works on extractive summarization use two major steps:

1) the first step is ranking the sentences based on their scores which are computed by combining features such as term frequency (TF), positional information and cue phrases;
2) the second step consists in selecting a few top ranked sentences to prepare the summary

Our work

- In this work we present an initial effort to develop a Hungarian speech summarization system.
- Summarization will also be compared to a baseline version using tokens available from human annotation.
- In current work we propose a prosody based automatic tokenizer which recovers intonational phrases (IP) and use IPs as sentence like units in further analysis.
- The baseline tokenization relies on acoustic (silence) and syntactic-semantic (syntactically or semantically closely together belonging) axes.
- In Hungarian, both speech recognition and text-based syntactical analysis are difficult compared to English due to the very rich morphology of the language.

MATERIAL AND SPEECH-TO-TEXT

SPEECH MATERIAL

- 4 interviews from the BEA Hungarian Spontaneous Speech database
- Participants talk about their jobs, family, and hobbies.
- Three of the speakers are male and one of them is female.
- All speakers are native Hungarian, living in Budapest (aged between 30 and 60).
- The total material is 28 minutes long (average duration was 7 minutes per participant).

The Speech-to-Text System

- We use 160 interviews from BEA, accounting for 120 hours of speech (the interviewer discarded) to train Speech-to-text (S2T) acoustic models.
- Speakers involved in the 4 interviews used for summarization are held out.
- Using the Kaldi toolkit we train 3 hidden layer DNN acoustic models with 2048 neurons per layer and tanh non-linearity on 160 interviews from BEA (Hungarian).
- Input data is 9x spliced MFCC13 + CMVN +LDA/MLLT.
- A trigram language model is trained on transcripts of the 160 interviews after text normalization, with Kneser-Ney smoothing. Dictionaries are obtained using a rule-based phonetizer (spoken Hungarian is very close to the written form).

Word Error Rate (WER) was found around **44%** for this task. This relative high WER is justified by the high spontaneity of speech.
Stem error rate was found to be somewhat smaller, **39%**.

UTTERANCE SEGMENTATION (THE IP-TOKENIZER)

- This system uses phonological phrase models and aligns them to the input speech based on prosodic features F0 and mean energy. (Szaszák and Beke 2012).
- In this work we use it to obtain sentence-like tokens from speech-to-text output.

We use the IP tokenizer in an operating point with high precision (**96%** on read speech) and lower recall (**80%** on read speech).

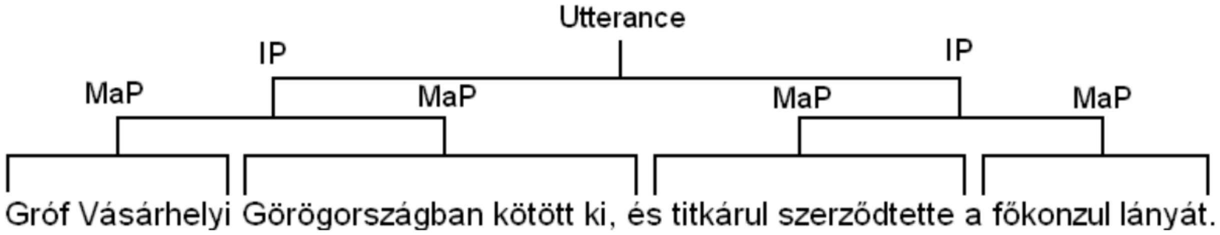
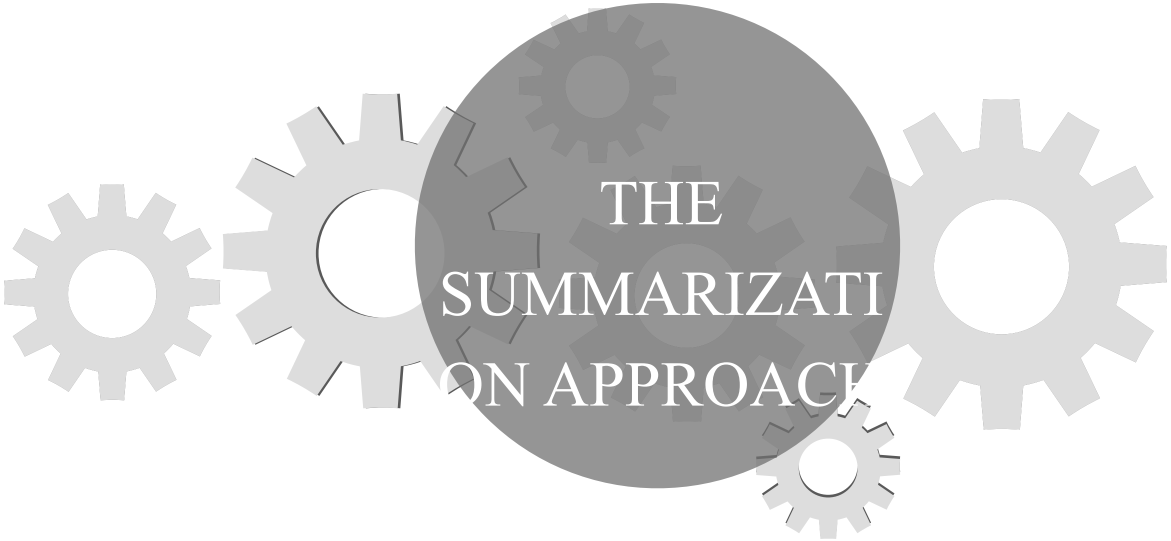
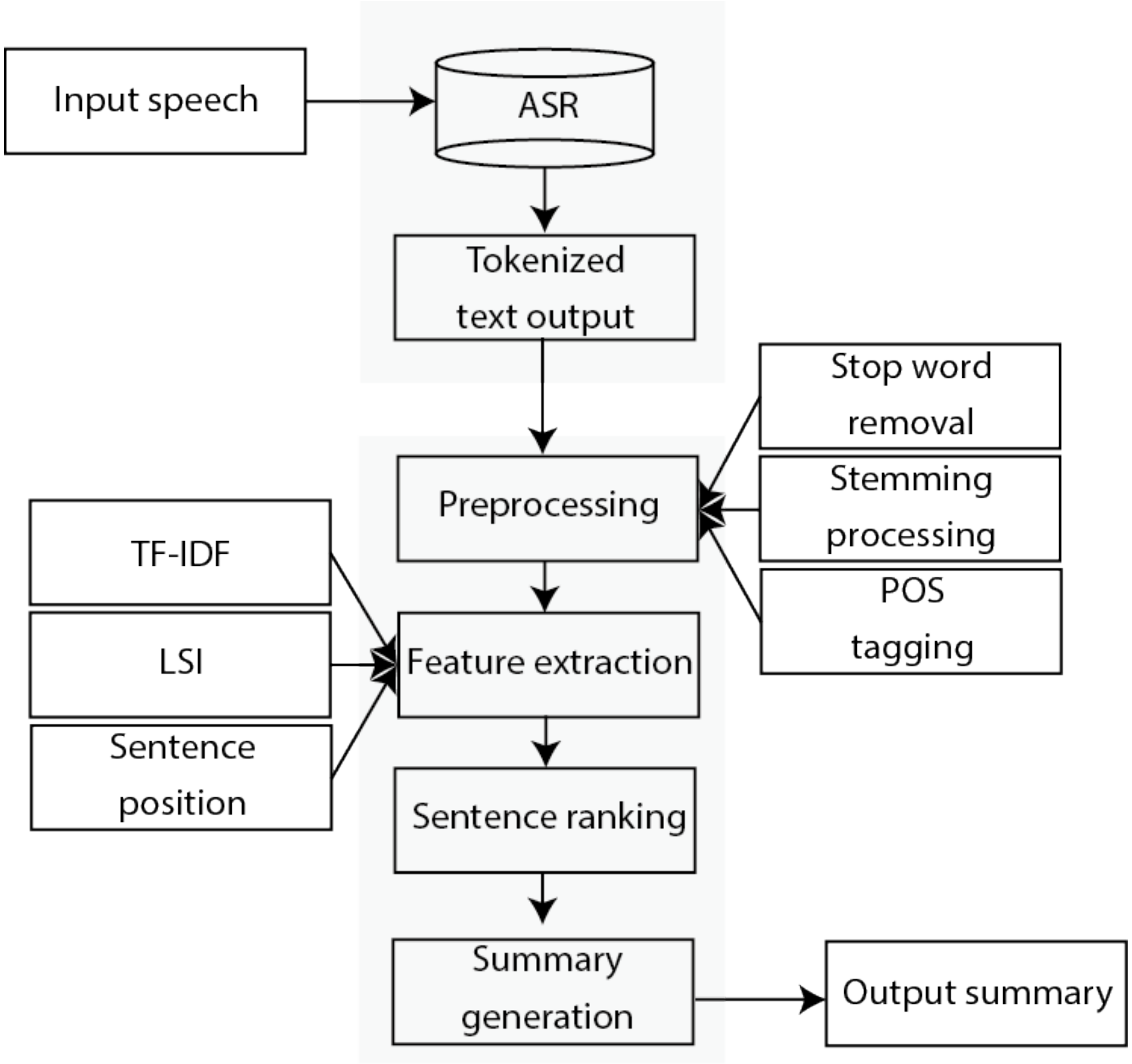

Utterance
IP
IP
MaP
MaP
MaP
MaP
Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát.

Figure 2: An example of canonical prosodic structure of a Hungarian sentence *"Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát"*



THE SUMMARIZATION APPROACH

BLOCK DIAGRAM



PRE-PROCESSING

- Stop words are removed from the tokens and stemming is performed. Stop-words are collected into a list, which contains
  - all words tagged as fillers by the S2T component (speaker noise) and
  - a predefined set of non-content words such as articles, conjunctions etc.
- The *magyarlánc toolkit* (Zsibrita et al. 2013) was used for the stemming and POS-tagging of the Hungarian text.
  - The words are filtered to keep only nouns.

TEXTUAL FEATURE EXTRACTION (WORD LEVEL)

- **TF-IDF (Term Frequency - Inverse Document Frequency)** reflects the importance of a sentence and is generally measured by the number of keywords present in it. The importance value of a sentence is computed as the sum of TF-IDF values of its constituent words (in this work: nouns) divided by the sum of all TF-IDF values found in the text.
- **Latent Semantic Analysis (LSA)** exploits context to try to find words with similar meaning. LSA is able to reflect both word and sentence importance. Singular Value Decomposition (SVD) is used to assess semantic similarity.

TEXTUAL FEATURE EXTRACTION (SENTENCE LEVEL)

- **Positional Value:** the more meaningful sentences can be found at the beginning of the document.
  - This is even more true in case of spontaneous narratives, as the interviewer asks the participant to tell something about her/his life, job, hobbies

$$P_k = 1/\sqrt{k}$$

where the P_k is the positional score of kth sentence.

- **Sentence length**: Usually a short sentence is less informative than a longer one and hence, readers or listeners are more prone to select a longer sentence than a short one when asked to find good summarizing sentences in documents. If a sentence is too short or too long, it is assigned a ranking score of 0.

- **Sentence ranking**: The ranking score RS_K is calculated as the linear combination of the so-called thematic term based score S_k and positional score P_k. The final score of a sentence k is:

$$RS_k = \begin{cases} \alpha S_k + \beta P_k, & if L_k \geq L_L \quad \& \quad L_k \leq L_U \\ 0 \quad otherwise, \end{cases}$$

where α is the lower, β is the upper cut-off for the sentence position ($0 \leq \alpha, \beta \leq 1$) and L_L is the lower and L_U is the upper cut-off on the sentence length L_k.

SUMMARY GENERATION

The last step is to generate the summary. In this process the N-top ranked sentences are selected from the text (Sarkar 2012).

We set N to 10, so the final text summary contains the top 10 sentences.

EVAULATION

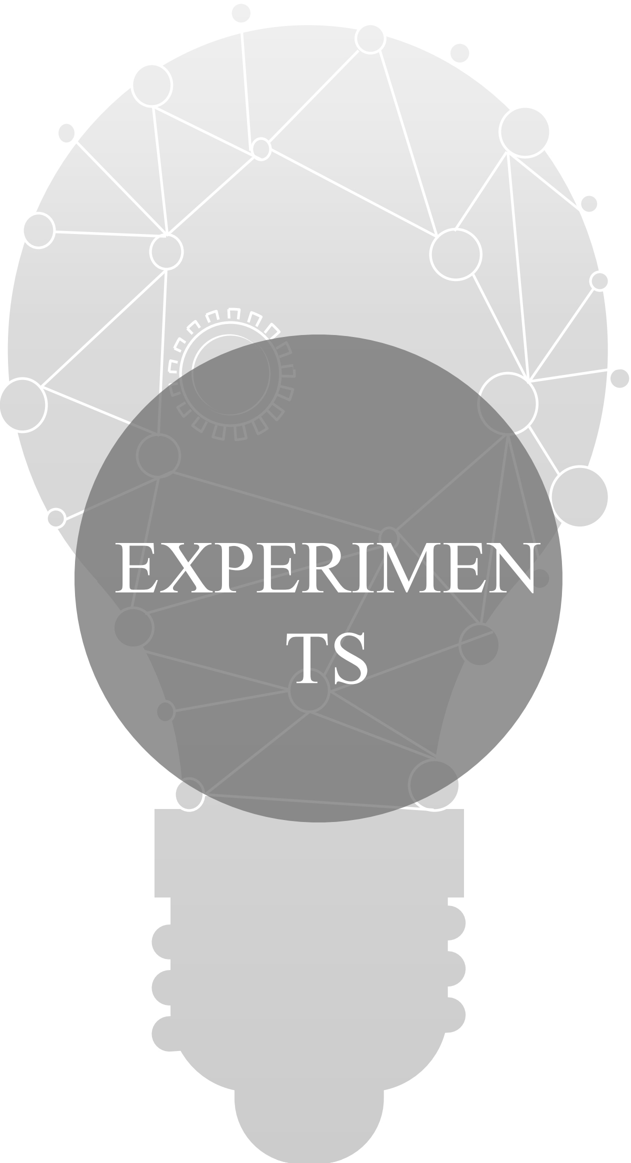
METRICS

Create reference summary

- 10 participants were asked to select up to 10 sentences that they find to be the most informative for a given document (presented also in spoken and in written form).
- Participants used 6.8 sentences on average for their summaries. For each narrative, a set of reference sentences was created: sentences chosen by at least 1/3 of the participants were added to the reference summary.

Compare reference summary and system summary

- **OBJECTIVE**
  - **F1-measure:** soft comparison
  - **ROUGE:** hard comparison



EXPERIMENTS

EXPERIMENST

3 setups:

- **OT-H**: Use the original transcribed text as segmented by the human annotators into sentence-like units.
- **S2T-H**: Use speech-to-text conversion to obtain text, but use the human annotated tokens.
- **S2T-IP**: Use speech-to-text conversion to obtain text and tokenize it based on IP boundary detection from speech.

	Soft comparison			Hard comparison			
Setup	Approach	Recall %	Precision %	F1	Recall	Precision	F1
OT-H	TF-IDF	0.51	0.76	0.61	0.36	0.28	0.32
	LSA	0.36	0.71	0.46	0.36	0.3	0.32
S2T-H	TF-IDF	0.51	0.8	0.61	0.34	0.29	0.31
	LSA	0.49	0.77	0.56	0.39	0.27	0.32
S2T-IP	TF-IDF	0.62	0.79	0.68	0.33	0.28	0.30
	LSA	0.59	0.78	0.65	0.33	0.32	0.32

CONCLUSI ON

CONCLUSION

- This paper addressed speech summarization for highly spontaneous Hungarian. Given this high degree of spontaneity and also the heavy agglutinating property of Hungarian, we beleive the obtanied results are promising as they are comparable to results published for other languages (Campr, M., Ježek 2015).
- The proposed IP detection based tokenization was as successful as the available human one. The overall best results were 62% recall and 79% precision (F1 = 0.68).

THANK U