Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts

Veronika Vincze[1,2], Gábor Gosztolya[1,2], László Tóth[1], Ildikó Hoffmann[2,3], Gréta Szatlóczki[2], Zoltán Bánréti[3], Magdolna Pákáski[2], János Kálmán[2]

[1] MTA-SZTE Research Group for Artificial Intelligence
[2] University of Szeged
[3] Research Institute of Linguistics, Hungarian Academy of Sciences
vinczev@inf.u-szeged.hu

Motivation and goals

- Alzheimer's disease (AD) may take years to develop
- in its early stage it usually appears as a mild cognitive impairment (MCI)
- it is very difficult to diagnose
- yet an early diagnosis would be important (to start the treatment as early as possible)
- Goal: to develop an automatic screening tool for MCI
- NOT a perfect diagnosis but a quick and cheap pre-filtering of the population

Mild cognitive impairment

- prodromal stage of Alzheimer's Disease
- the (spontaneous) speech of the patient is influenced
- verbal fluency declines: longer hesitations and a lower speech rate
- the lexical frequency of words and part-of-speech tags may also change
- the emotional responsiveness of the patient also changes
- however, language capacities have received marginal attention when diagnosing AD [1]

Automatic detection of dementia

- automatic speech recognition tools for detecting
  - aphasia [2]
  - mild cognitive impairment [3]
  - Alzheimer's Disease [4]
- lexical analysis of spontaneous speech [5]
- changes in the writing style may also refer to dementia [6]
- speech recognition techniques for detecting MCI in Hungarian [7]

ours is the first attempt to identify MCI on the basis of written texts for Hungarian

Data collection

- 84 native speakers of Hungarian (a morphologically rich language), whose medical diagnosis for MCI were at our disposal
- two short animated films were presented to the patients at the memory ambulance of the University of Szeged
- patients were asked to talk about the first film then about their previous day, and lastly, about the second film
- speech productions were recorded and transcribed by linguists

	MCI	Control	Total
Male	16	13	29
Female	32	23	55
Total	48	36	84

collaboration of linguists, medical experts and computer scientists

Linguistic features of transcripts

- several forms of hesitations and silent pauses
- phonological deletion (*mer* instead of the standard form *mert* „because")
- lengthening (*utánna* instead of the standard form *utána* „then")
- duplications (*ez ezt* „this this-ACC")
- neologisms (*feltkáva*, probably *főtt kávé* „boiled coffee")
- fillers, indefinite pronouns and uncertain words (*ilyen* „such", *izé* „thing, gadget", *és aztán* „and then", *valamilyen* „some kind of", *valahogy* „somehow")
- paraphrases (*egy ilyen bagolyszerűség* a such owl-likeness „something similar to an owl")

Experiments

- transcripts were morphologically and syntactically parsed by magyarlanc [8]
- morphological, syntactic and semantic features were extracted from the output of magyarlanc
- statistically significant differences were found for most of the features
- support vector machines with leave-one-out cross validation
- baseline: majority labeling (57.14% in terms of accuracy)

Feature set

Spontaneous speech based features: filled and silent pauses; hesitations; pauses that follow an article and precede content words; lengthened sounds (as a special form of hesitation)

Morphological features: number of tokens and words; number and rate of distinct lemmas; number of punctuation marks; number and rate of nouns, verbs, adjectives, pronouns and conjunctions; number of first person singular verbs; number and rate of unanalyzed words

Semantic features: fillers and uncertain words compared to the number of all tokens; words/phrases related to memory activity (e.g. *nem emlékszem* not remember-1SG „I can't remember"); negation words; content words and function words; number of thematic words related to the content of the films

Demographic features: gender; age; education.

Results

	MCI			Control			Total			
Features	P	R	F	P	R	F	P	R	F	%
all included	72.0	75.0	73.5	64.7	61.1	62.9	68.9	69.0	68.9	69.1
w/o semantic	75.0	81.3	78.0	71.9	63.9	67.6	73.7	73.8	73.6	73.8
	+3.0	+6.3	+4.5	+7.2	+2.8	+4.7	+4.8	+4.8	+4.7	+4.7
w/o demographic	70.0	72.9	71.4	61.8	58.3	60.0	66.5	66.7	66.5	66.7
	-2.0	-2.1	-2.1	-2.9	-2.8	-2.9	-2.4	-2.3	-2.4	-2.4
w/o speech-based	70.8	70.8	70.8	61.1	61.1	61.1	66.7	66.7	66.7	66.7
	-1.2	-4.2	-2.7	-3.6	0.0	-1.8	-2.2	-2.3	-2.2	-2.4
w/o morphological	72.3	70.8	71.6	62.2	63.9	63.0	68.0	67.9	67.9	67.9
	+0.3	-4.2	-1.9	-2.5	+2.8	+0.1	-0.9	-1.1	-1.0	-1.2
only significant	81.4	72.9	76.9	68.3	77.8	72.7	75.8	75.0	75.1	75.0
	+9.4	-2.1	+3.4	+3.6	+16.7	+9.8	+6.9	+6.0	+6.2	+5.9

Discussion

- statistically significant differences among MCI patients and healthy controls concerning several linguistic and speech-based features
- speech-based, demographic and morphological features unequivocally contributed to performance
- the effect of semantic features seems less obvious as they harm performance taken as a whole but some individual semantic features are useful for the system
- MCI patients that spoke only a few short sentences were often classified as healthy controls due to a lower number and rate of hesitations, pauses, fillers and uncertain words
- healthy subjects who talked more also hesitated more, hence they were misclassified as MCI patients

Conclusions

- automatic detection of Hungarian patients suffering from mild cognitive impairment on the basis of their speech transcripts
- both statistical and machine learning results revealed that morphological and spontaneous speech-based features have an essential role in distinguishing MCI patients from healthy controls
- Future work:
  - dataset to be expanded
  - machine learning system to be improved by combining features from automatic speech recognition and from the analysis of written texts

References

[1] Bayles, K.A.: Language function in senile dementia. Brain and Language **16**(2) (1982) 265–280

[2] Fraser, K.C., Meltzer, J.A., Graham, N.L., Leonard, C., Hirst, G., Black, S.E., Rochon, E.: Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex **55** (2014) 43–60

[3] Lehr, M., Prud'hommeaux, E.T., Shafran, I., Roark, B.: Fully automated neuropsychological assessment for detecting mild cognitive impairment. In: INTERSPEECH, ISCA (2012) 1039–1042

[4] Satt, A., Hoory, R., König, A., Aalten, P., Robert, P.H.: Speech-based automatic and robust detection of very early dementia. In: 15th Annual Conference of the International Speech Communication Association. (2014) 2538–2542

[5] Thomas, C., Keselj, V., Cercone, N., Rockwood, K., Asp, E.: Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In: Mechatronics and Automation, 2005 IEEE International Conference. Volume 3., IEEE (2005) 1569–1574

[6] Le, X., Lancashire, I., Hirst, G., Jokel, R.: Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Literary and Linguistic Computing **26**(4) (2011) 435–461

[7] Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., Kálmán, J.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: 16th Annual Conference of the International Speech Communication Association. (2015) 2694–2698

[8] Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771