Linguistic Calculations on Cyfronet High Performance Computers

Bartosz Ziółko, Mariusz Ziółko
Department of Electronics
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059, Kraków, Poland
{bziolko,ziolko}@agh.edu.pl

Introduction to our research

The Signal Processing Group at AGH University of Science uses the Cyfronet high performance computers to process linguistic data to construct the Polish language models. The results will be applied to an large vocabulary automatic speech recognition system. Natural language processing always faces problems of data sparsity. The quality of language models depends strongly on the amount of text corpora available during the training. This is why, there is a trade-off of quality and time spent on calculations. The high performance computers facilitate obtaining the linguistic rules from the huge amount of texts written in Polish.

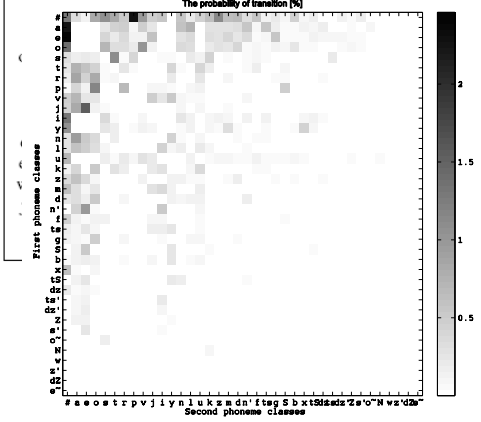
The aim of our experiment

There are several calculation tasks which we conduct or plan to conduct on computers Mars and Saturn. Currently, the main one is finding triphone statistics for Polish language. Our first attempt to this task was already published [1]. The task was conducted on the Parliament transcriptions mainly, which is around 50 megabytes of text. Now, we want to repeat the process for data of around 2 gigabytes. Such statistics are very useful in speech modelling.

It is very difficult to provide the proper acoustic data for all possible triphones, but there are methods to synthesise no-recorded ones using data for other triphones and phonological similarities between different phonemes. It means that the list of possible triphones has to be provided for a given language. The triphone statistics can be also used to generate hypotheses used in recognition of out-of-dictionary words.

Obtaining of phonetic information from an orthographic text-data is not straightforward. Transcription of text into phonetic data has to be applied first. We used PolPhone [2] software for this aim. The SAMPA extended phonetic alphabet was applied with 39 symbols and pronunciation rules typical for cities Kraków and Poznań. We altered the PolPhone phonetic alphabet to a 37 symbol version which is used in the largest corpus

SAMPA	example	transcr.	occurr.	%
#		#	23,810,956	16.086,7
a	pat	pat	13,311,163	8.993
e	test	test	11,871,405	8.020,3
o	pot	pot	10,566,010	7.138,4
s	syk	sIk	5,716,058	3.861,8
t	test	test	5,703,429	3.853,2
r	ryk	rIk	5,171,698	3.494
p	pik	pik	5,150,964	3.48
v	wilk	vilk	5,025,050	3.394,9
j	jak	jak	4,996,475	3.375,6
i	PIT	pit	4,994,743	3.374,4
I	typ	tIp	4,974,567	3.360,8
n	nasz	naS	4,602,314	3.109,3
l	luk	luk	4,399,366	2.972,2
u	puk	puk	4,355,825	2.942,8
k	kit	kitk	4,020,161	2.716
z	zbir	zbir	3,602,857	2.434,1
m	mysz	mIS	3,525,813	2.382
d	dym	dIm	3,267,009	2.207,2
n'	koń	kon'	3,182,940	2.150,4
f	fan	fan	2,030,717	1.372
ts	cyk	tsIk	1,984,311	1.340,6
g	gen	gen	1,949,890	1.317,3
S	szyk	SIk	1,739,146	1.175
b	bit	bit	1,668,103	1.127
x	hymn	xImn	1,339,311	0.904,84
tS	czyn	tSIn	1,285,310	0.868,36
dz	dzwoń	dzvon'	692,334	0.467,74
ts'	ćma	ts'ma	690,294	0.466,36
dz'	dźwig	dz'vik	589,266	0.398,11



of spoken Polish and currently recognised as a SAMPA standard. For programming reasons we used our own single letter only symbols corresponding to SAMPA symbols instead of typical ones to distinguish phonemes easier while analysing received phonetic transcriptions. Statistics can be now simply calculated by counting number of occurrences of each phoneme, phoneme pair, and phoneme triple in analysed text, where each phoneme is just one symbol.

Usage of Mars and Saturn for Linguistic Calculations

We use Matlab to analyse the text corpora. It is available on Mars and Saturn. We have more than 2 GigaBytes of data, as it is described in the table below and we still collect text corpora.

Content	MBytes
Parliament transcripts	58
Literature	490
Rzeczpospolita journal	879
Wikipedia	754

We found out some practical details about both computers. Mars can have only one session at the moment. We run a few processes in the background to analyse different parts of corpora pararerly. It was also necessary to split our text files because Matlab is not able to work on files larger then 98 megabytes. Saturn is able to have several sessions for the same user but works much slower than Mars. Starting 5 processes slows them down by around half, which is still more time efficient in total. The processes can be started by

nohup matlab < script.m &

The Matlab script.m file cannot have any input arguments, which is normally possible in Matlab if script.m is a function. Unfortunetlly, it does not work with *nohup*. The way to overcome it, is to put Matlab script with input arguments as a body of another script and run it. In example script.m can be simply

Func(arg1,arg2);.

It might be worth to mention that a Saturn session with processes in the background only was disconnected after some time. It does not happen with Mars. However, *nohup* seems to not work properly on Mars. It happens that the command is accepted, but the processes stop when the session is disconnected.

The content of the presentation and the paper

We aim to present lingustic data which are difficult to collect and are very useful for several researchers. We want to compare the results we gained using a regular PC with statistics obtained with help of high performance Cyfronet computers. We will present technical information about our experimental work was arranged, times of calculations and practical aspects of using Matlab on Mars and Saturn for processing large text files.

References:

[1] B. Ziółko, J. Gałka, S. Manandhar, R. C. Wilson, M. Ziółko, „Triphone Statistics for Polish Language”, *Proceedings of 3rd Language and Technology Conference*, Poznań, 2007.

[2] Demenko, G., M. Wypych, and E. Baranowska, „Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis”, *Speech and Language Technology, PTFon,* Poznań, 7(17), 2003.