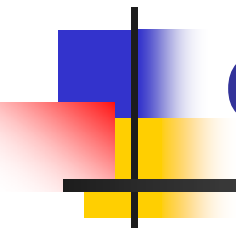
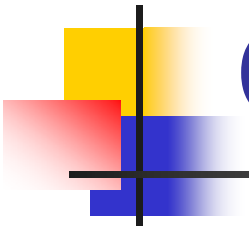


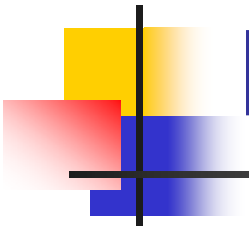
Natural languages and evaluation of their models

Tomáš Mikolov, 2007



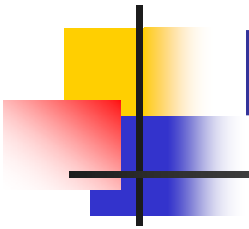
Overview

- Natural signals – what are we doing?
- History – who was working on this before?
- State of the art – how "good" are the present methods?
- Future
- Evaluation: speech recognition



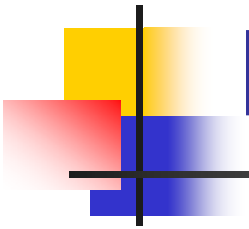
Natural signals

- Anything that can anyone observe in the universe
- Related to set of languages accepted by Turing machines



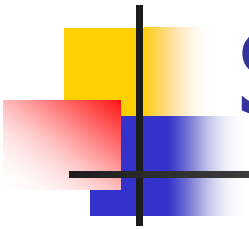
Natural signals

- Our aim is to describe these signals in the best possible way
- If we would be able to do it, we can solve any problem



History

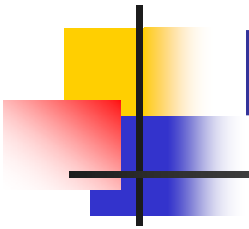
- Turing – definition of Turing machine, capable of computing anything that is computable
- Shannon – information theory (transmission of information over noisy channel), definition of entropy
- Kolmogorov (and others) – definition of Kolmogorov complexity
- Solomonoff – algorithmic probability



State of the art

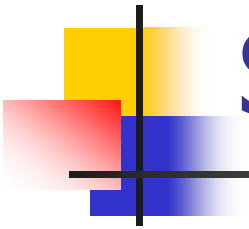
- Markov models (N-grams, HMM, NN, …) - probabilistic finite state automatons
- PCFGs – a lot of work, but not widely used in practice

- Applications in speech recognition, machine translation, computer vision, …



Future - need better models

- FSA are clearly incapable of learning effectively some relationships (long context dependencies due to no recursion)
- CFGs are computationally expensive and still very limited!
- Models with power of Turing machine?

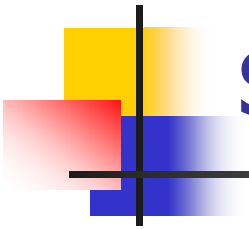


Speech recognition

- Acoustic models
- Language models

- Sample output:
- SEKVENCÍ ČÍSEL TAKY BUDEME NĚKDY ŘÍKAT POSLOUPNOSTI PODOBNĚ A K TOMU ŘÍKAJÍ MATEMATICI JE TO V ZÁSADĚ ÚPLNĚ JEDNO JAK SE TOMU ŘÍKÁ PROSTĚ
- PROTOŽE TEN DISKRÉTNÍ JEDNOTKOVÝ IMPULZ JE VOPRAVDU NĚCO CO SE DÁ PLNĚ BEZPROBLÉMŮ VYGENEROVAT JÁ SE S TÍM POČÍTAT JE TO JEDNIČKA NA VZORKU NULA A NULY VŠUDE JINDE

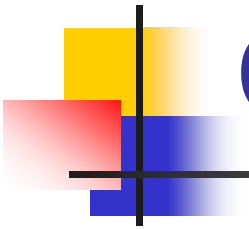
Evaluation of different models: speech recognition



- The easiest way how to evaluate new models is to rescore N-best lists (list of possible hypotheses with acoustic and language score)
- More intelligent models provide more accurate results (for example, neural networks are better than plain n-grams)

Example of N-best list

```
-1890.09 -5.45964 2 <s> <s> PĚT MINUT </s> </s>
-1904.73 -7.00808 2 <s> <s> JE MINUT </s> </s>
-1909.26 -7.22184 2 <s> <s> TY MINUT </s> </s>
-1910.83 -7.10182 3 <s> <s> PĚT MINUT A </s> </s>
-1904.59 -7.67013 2 <s> <s> TEĎ MINUT </s> </s>
-1906.21 -7.78719 3 <s> <s> JE TO MINUT </s> </s>
-1894.68 -8.96496 3 <s> <s> PĚT MINUT V </s> </s>
-1914.19 -7.35332 2 <s> <s> JEDNY NO </s> </s>
-1891.93 -9.29704 2 <s> <s> JET MINUT </s> </s>
-1896.18 -8.9921 3 <s> <s> PĚT MINUT S </s> </s>
-1896.04 -9.01288 3 <s> <s> PĚT MINUT Z </s> </s>
-1888.03 -9.81826 2 <s> <s> JEDU MINUT </s> </s>
-1918.76 -7.37834 2 <s> <s> PĚT MÍNUS </s> </s>
```



Conclusion

- If anyone is interested in natural language processing, I can provide data for experiments