Learning Simple Phonotactics

Erik F. Tjong Kim Sang
Center for Dutch Language and Speech
University of Antwerp
Belgium
erikt@uia.ua.ac.be

John Nerbonne
Alfa-informatica, BCN
University of Groningen
The Netherlands
nerbonne@let.rug.nl

overhead sheets: http://lcg-www.uia.ac.be/~erikt/talks/

Goal

Examine application of machine learning algorithms for deriving natural language knowledge.

Task

Deriving models for the phonotactic structure of monosyllabic Dutch words:

bad	$\mapsto$	GOOD
crwth	$\mapsto$	BAD
mlod	$\mapsto$	BAD
teen	$\mapsto$	GOOD
zachtst	$\mapsto$	GOOD

Research Questions

1. What machine learning technique generates the best models?
2. What is the influence of knowledge representation on the performances?
3. Does initial language knowledge improve the results?

Results orthographic data representation

	non-initialized		initialized	
	positive	negative	positive	negative
HMM	98.9%	91.0%	98.9%	94.5%
ILP	99.2%	60.0%	97.8%	97.7%
SRN	100.0%	8.3%	100.0%	4.8%
baseline	99.2%	60.2%		

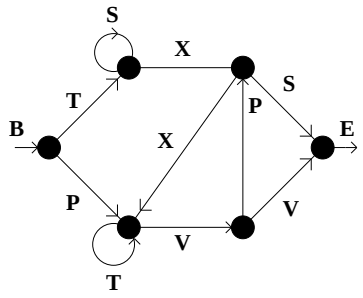
Results phonetic data representation

	non-initialized		initialized	
	positive	negative	positive	negative
HMM	99.1%	98.3%	99.1%	99.1%
ILP	99.2%	70.6%	99.2%	98.3%
baseline	99.2%	71.6%		

Note: these results come from [Tjo98]. The results in the paper are 10-cv results for the same data set.

SRN performance explanation

[Cle93]: SRNs are able to recognize Reber grammar strings with 100% accuracy.



However, the performance will degrade when the complexity of the grammar is increased:

complexity	positive	negative
2	100.0%	100.0%
3	100.0%	92.2%
4	100.0%	82.2%

Cause: signal/noise ration in network goes down when the grammar becomes more complex.

Concluding remarks

1. What machine learning technique generates the best models?

↦ HMMs and ILP produce good models; SRNs don't.

2. What is the influence of knowledge representation on the performances?

↦ The models generated for phonetic data are nearly always better than the models generated for orthographic data.

3. Does initial language knowledge improve the results?

↦ Yes, the availability of initial linguistic knowledge improves the performance, especially in the case of ILP.

Address

Address:	Center for Dutch Language and Speech University of Antwerp Universiteitsplein 1 B-2610 Wilrijk Belgium
telephone:	+31 3 8202765
WWW:	http://lcg-www.uia.ac.be/~erikt/

References

[Cle93] Axel Cleeremans. *Mechanisms of Implicit Learning*, The MIT Press, 1993. ISBN 0-262-03205-8.

[Tjo98] Erik F. Tjong Kim Sang, *Machine Learning of Phonotactics*, PhD thesis University of Groningen, The Netherlands, 1998. http://lcg-www.uia.ac.be/~erikt/mlp/

Overhead sheets

http://lcg-www.uia.ac.be/~erikt/talks/