UPPSALA
UNIVERSITET

Big is Beautiful or Less is More?

Reflections on Resource-Intensive NLP

Joakim Nivre

Uppsala University
Department of Linguistics and Philology

Introduction

Introduction

Natural language processing

- Make computers do useful and interesting things with language
- Gain insights about human language from computational models

Introduction

Natural language processing

- Make computers do useful and interesting things with language
- Gain insights about human language from computational models

From armchair linguistics:

- Carefully hand-crafted grammars with limited coverage – toy systems

Introduction

Natural language processing

- Make computers do useful and interesting things with language
- Gain insights about human language from computational models

From armchair linguistics:

- Carefully hand-crafted grammars with limited coverage – toy systems

To big data:

- Broad-coverage statistical systems trained on large amounts of data

The Google Effect

The Google Effect

State of the art in many areas of NLP requires:

- Massive amounts of data
- High-performance computing

The Google Effect

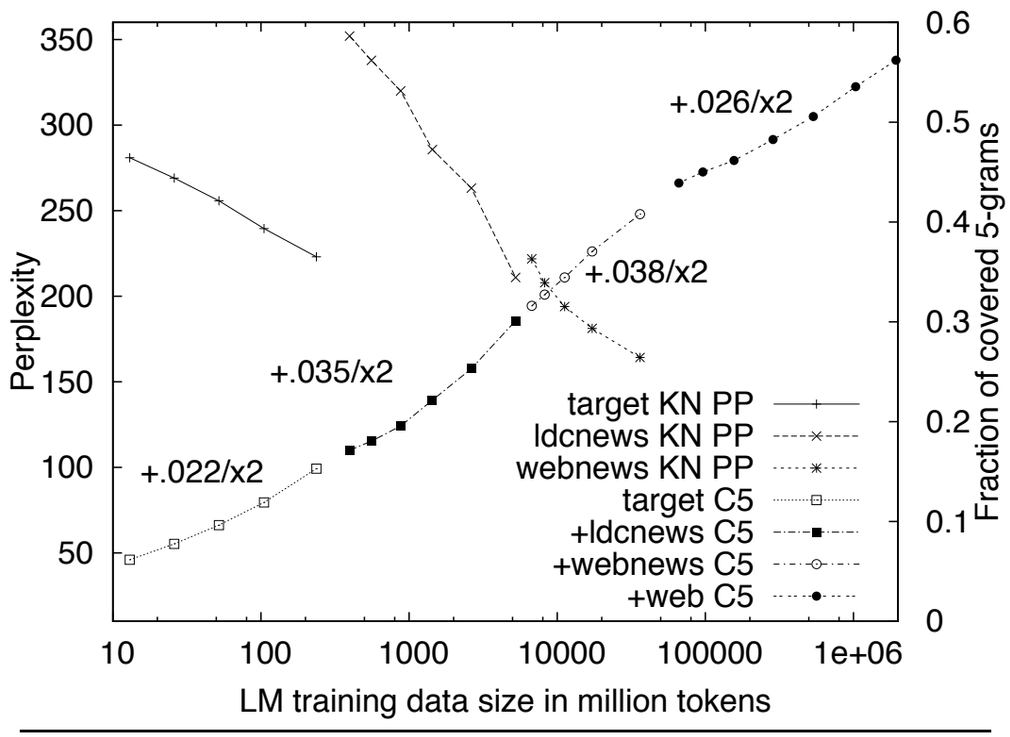
State of the art in many areas of NLP requires:

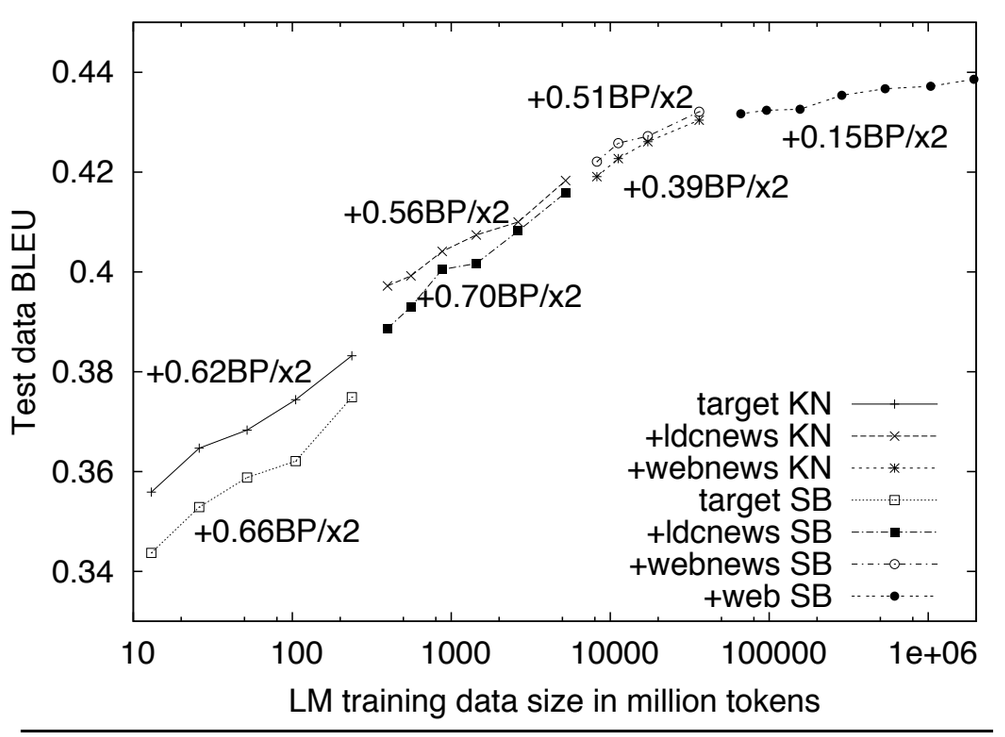
- Massive amounts of data
- High-performance computing

Emerging trend:

- Academic institutions can't keep up with industrial labs

Machine Translation





Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och and Jeffrey Dean. 2007. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 858–867.

Syntactic Parsing

Type	System	UAS	COMP
Sup	McDonald06	91.5	
	Koo10	93.04	-
	Zhang11	92.9	48.0
	Li12	93.12	-
	Our Baseline	92.76	48.05
Semi	Koo08	93.16	
	Suzuki09	93.79	
	Chen09	93.16	47.15
	Zhou11	92.64	46.61
	Suzuki11	94.22	-
	Chen12	92.76	-
	MetaParser	93.77	51.36

Wenlian Cheng, Min Zhang and Yue Zhang.. 2013. Semi-Supervised Feature Transformation for Dependency Parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1304–1313.

Syntactic Parsing

Type	System	UAS	COMP
Sup	McDonald06	91.5	
	Koo10	93.04	-
	Zhang11	92.9	
	Li12	93.12	-
	Our Baseline	92.76	48.05
Semi	Koo08	93.16	
	Suzuki09	93.79	
	Chen09	93.16	47.15
	Zhou11	92.64	[illegible]
	Suzuki11	94.22	-
	Chen12	92.76	-
	MetaParser	93.77	51.36

Small Data

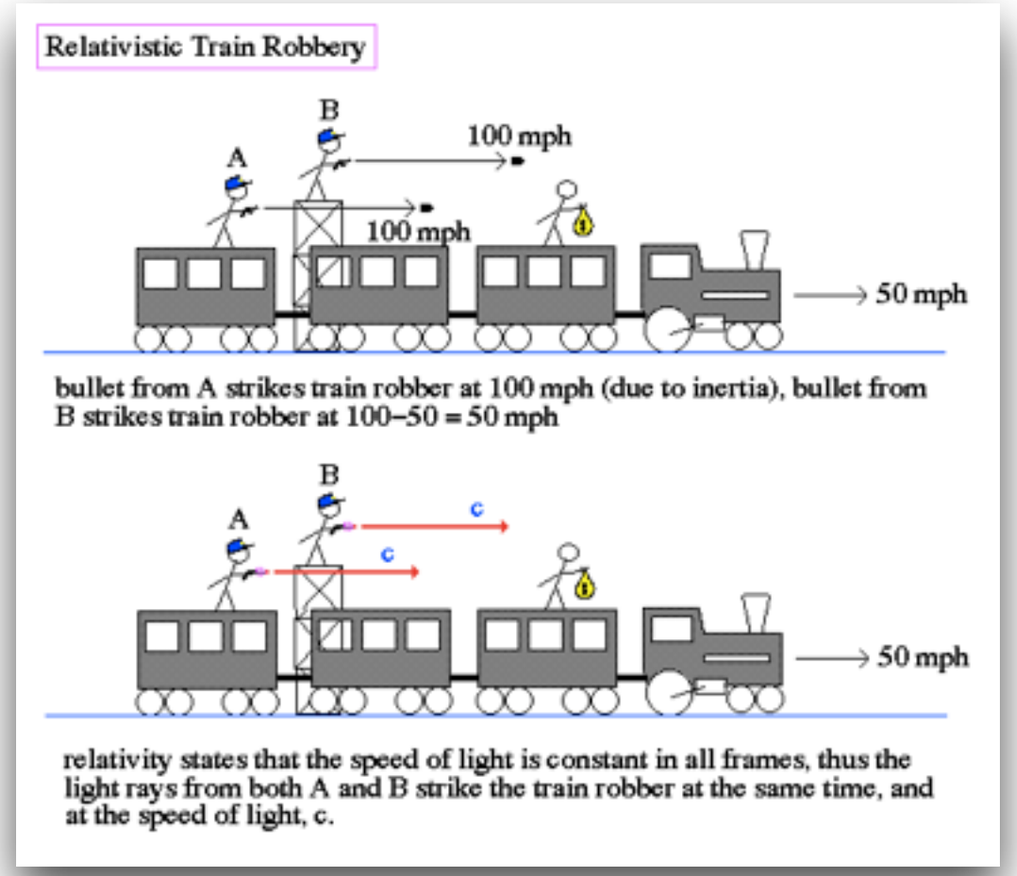
Big Data

Wenlian Cheng, Min Zhang and Yue Zhang.. 2013. Semi-Supervised Feature Transformation for Dependency Parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1304–1313.

A scientific issue?

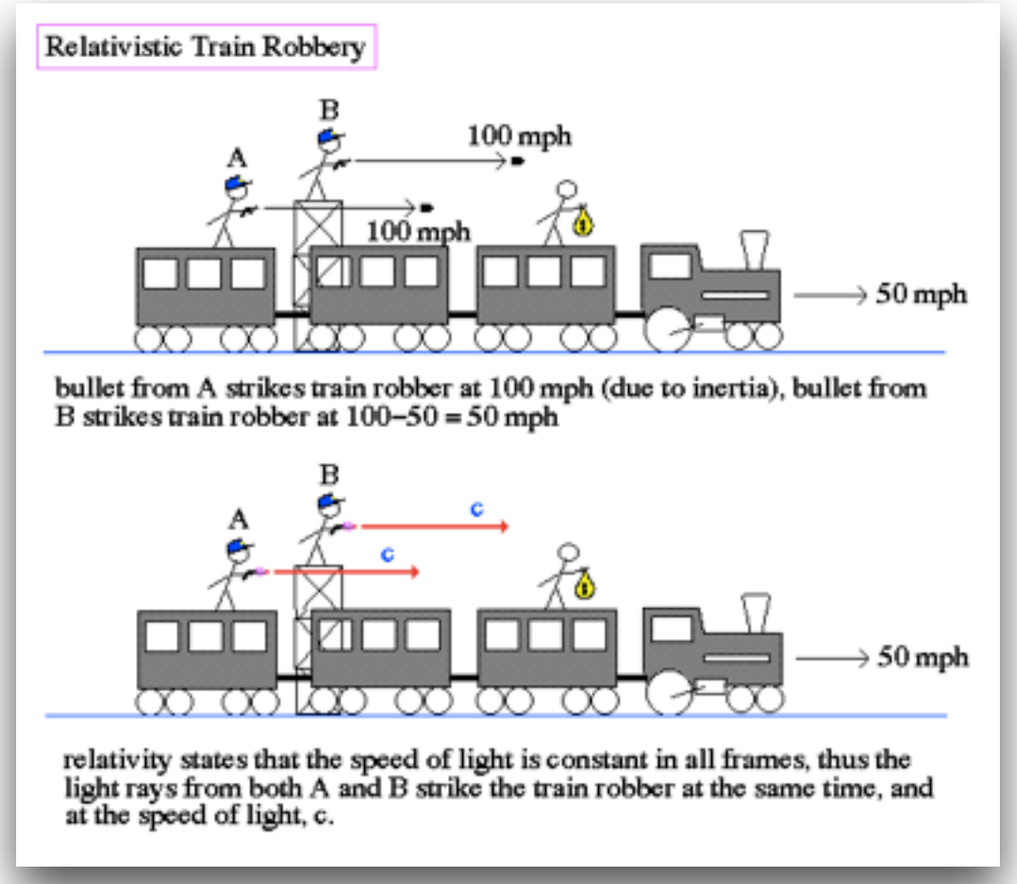
A scientific issue?

Galilean vs. Einsteinian Relativity

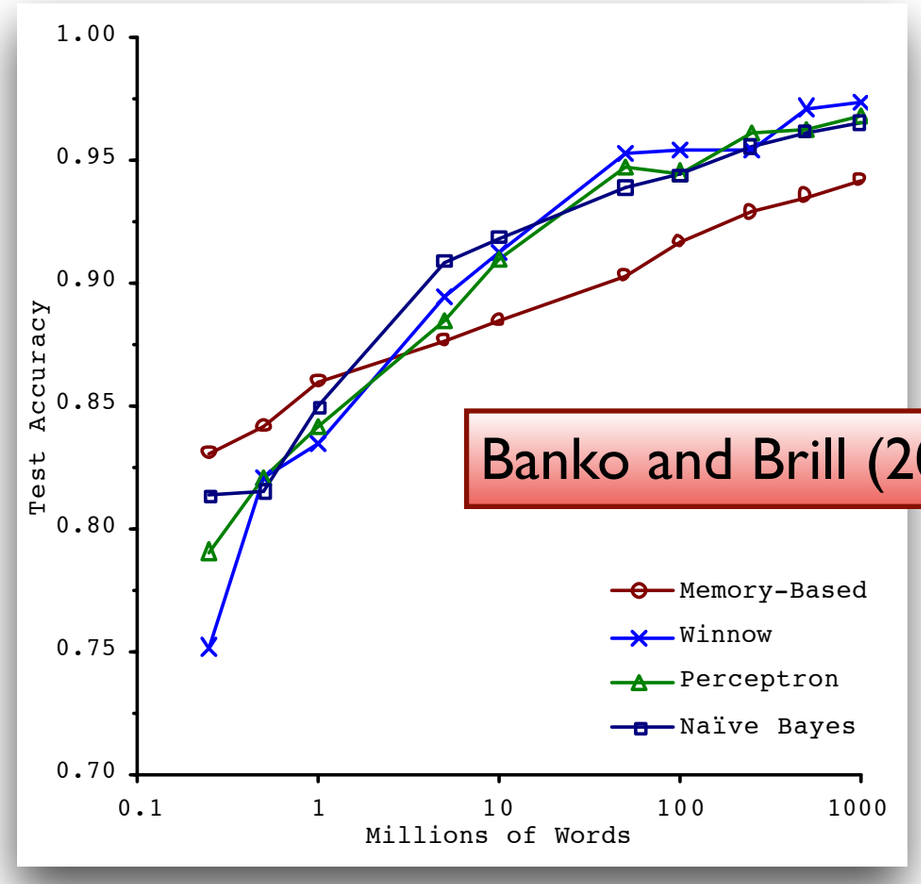


A scientific issue?

Galilean vs. Einsteinian Relativity



Small vs. Big Data



Banko and Brill (2001)

What can we do?

What can we do?

Go to work in industrial labs?

- But then all NLP research will be driven by industry goals
- And who will educate their researchers and engineers in the future?

What can we do?

Go to work in industrial labs?

- But then all NLP research will be driven by industry goals
- And who will educate their researchers and engineers in the future?

Create our own niche?

- Machine translation for extinct languages (small data)
- Parsers trained on only treebank data and run on a single processor

What can we do?

Go to work in industrial labs?

- But then all NLP research will be driven by industry goals
- And who will educate their researchers and engineers in the future?

Create our own niche?

- Machine translation for extinct languages (small data)
- Parsers trained on only treebank data and run on a single processor

What do we need?

What do we need?

Scientific requirements:

- Hypothesis testing under relevant conditions
- Training and testing on the whole range of data set sizes
- Slow train NLP as well as NLP near the speed of light

What do we need?

Scientific requirements:

- Hypothesis testing under relevant conditions
- Training and testing on the whole range of data set sizes
- Slow train NLP as well as NLP near the speed of light

Infrastructure requirements:

- Massive amounts of data
- High-performance computing
- Expertise on how to use these resources
- Calls for a community effort

Reminder

Reminder

Why is Google using 10-year old parsing technology?

- The scientific literature is full of more accurate parsers
- But they don't process 10,000 sentences per second

Reminder

Why is Google using 10-year old parsing technology?

- The scientific literature is full of more accurate parsers
- But they don't process 10,000 sentences per second

Sustainable NLP

- More data and computing power isn't always the answer
- We also need research on faster algorithms and leaner models

Conclusion

Conclusion

Big is beautiful

- The scientific community needs access to web-scale resources
- Otherwise much of our research will soon be irrelevant

Conclusion

Big is beautiful

- The scientific community needs access to web-scale resources
- Otherwise much of our research will soon be irrelevant

Less is more

- We are tackling infinite spaces with finite resources
- Making better use of resources is always going to be relevant

UPPSALA
UNIVERSITET

Thanks for Your Attention!

Questions?

http://stp.lingfil.uu.se/~nivre/