DATA DRIVEN MACHINE TRANSLATION

David Talbot

Spring 2017

Yandex School of Data Analysis

COURSE OVERVIEW

WHAT YOU WILL LEARN

Core problems of machine translation:

- Why machine translation is hard
- How to build a machine translation system from data
- How to evaluate a machine translation system
- How to improve it using linguistic knowledge

WHAT YOU WILL LEARN

Some of the topics we’ll focus on:

- Phrase-based machine translation
    - Word alignment
    - Syntax based reordering
    - Language models
- Neural machine translation
    - Word embeddings
    - Encoder-decoder models
    - Attention mechanisms
    - Challenges for NMT

what else you'll learn (beyond mt)

Practical aspects of statistical modelling

- Estimating a statistical model from data
- Handling hidden variables
- Engineering features in statistical models
- Regularizing models
- Addressing computational complexity in neural networks

HOW IT WILL BE ASSESSED

Open-ended assignments:

1. Word alignment (25%)
2. Tree-based reordering (25%)
3. Improving a baseline NMT system (25%)

In class participation and preparation (25%)

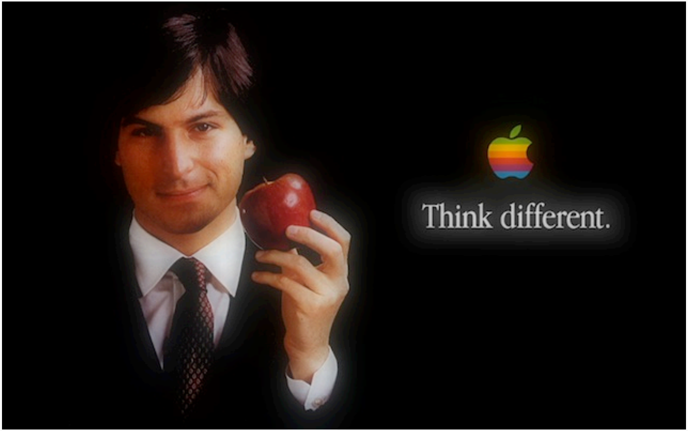
HOW WE WILL WORK

Lectures followed by practical sessions

- Lectures will introduce the main ideas
- Seminars will allow you to apply the ideas
- 25% credit will be given for active participation

WHY TRANSLATION IS HARD

AMBIGUITY



”Finally a computer that understands you like your mother.”
(Apple, 1985)

AMBIGUITY

”Finally a computer that understands you like your mother.”

AMBIGUITY

”Finally a computer that understands you like your mother.”

- В конце концов компьютер, который понимает вас так же хорошо, как ваша мама (понимает вас).

AMBIGUITY

”Finally a computer that understands you like your mother.”

- В конце концов компьютер, который понимает вас так же хорошо, как ваша мама (понимает вас).
- В конце концов компьютер, который понимает, что вам нравится ваша мама.

AMBIGUITY

”Finally a computer that understands you like your mother.”

- В конце концов компьютер, который понимает вас так же хорошо, как ваша мама (понимает вас).
- В конце концов компьютер, который понимает, что вам нравится ваша мама.
- В конце концов компьютер, который понимает вас так же хорошо, как он понимает вашу маму.

WORD ORDER

“As English not all languages words in the same order put. Hmmmmmm.” – Yoda

WORD ORDER

Languages can use very different word order

- He went to school by train.
- 彼は電車で学校に行きました。
- kara wa densha de gakkou ni ikimashita.

Why is reordering a huge problem for MT systems?

WORD ORDER

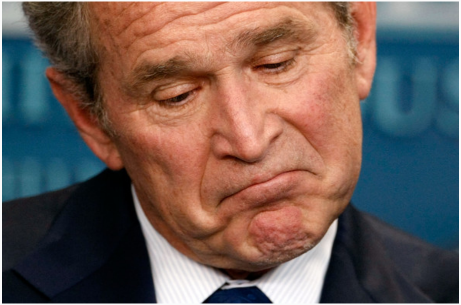
Languages can use very different word order

- He went to school by train.
- 彼は電車で学校に行きました。
- kara wa densha de gakkou ni ikimashita.

Why is reordering a problem for an MT system?

- A priori set of possible ’reorderings’ scales exponentially with sentence length

MORPHOLOGY



“Rarely is the question asked: Is our children learning? “
– George W. Bush

MORPHOLOGY

Many languages require case marking and agreement

- Перечень различных рекордов скорости, установленных на рельсовых путях, был ...
- Word choices depend on the gender, case, number etc. of other words
- Morphological agreement can span many tokens

INFORMATION STRUCTURE

Morphological case and agreement make word order less important

- The dog bit the hippopotamus.
- The hippopotamus bit the dog.

- Собака укусила бегемота.
- Собаку укусил бегемот.

WORLD KNOWLEDGE

Usually only one interpretation is reasonable for us

- Stolen painting found by tree.
- I haven’t slept for ten days.
- I saw a man with a telescope.

EVEN *EASY* SENTENCES ARE HARD

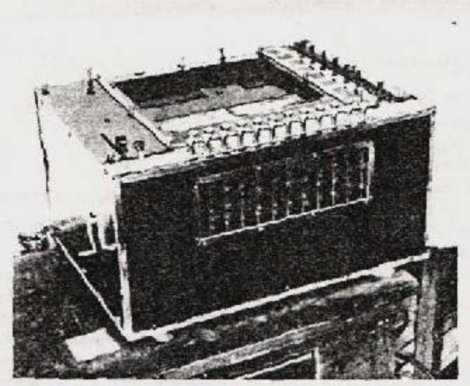
en	I’ve got two brothers.
fr	J’ai deux frères. (I’ve two brothers.)
ru	У меня два брата. (At me [are] two brothers.)
ja	私は2人の弟がいます. (As for me, two people younger brother there are.)

NOT A SIMPLE MACHINE LEARNING PROBLEM

- High dimensional: vocabulary > 1 million words
- Sparse: natural language follows a Zipf law
- Combinatorial: reordering is *a priori* $O(N!)$
- Dependencies
- Partially observed data
- More than one right answer...

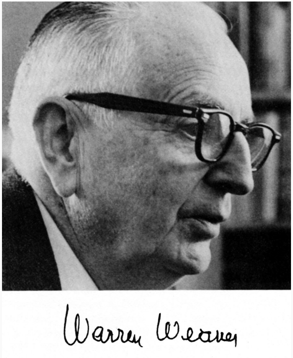
SOME OF THE STORY SO FAR

PRE-HISTORY



1930s Peter Troyanskii and Georges Artsrouni patented mechanical translation devices.

WW2 CRYPTOGRAPHY



1940s Shannaon, Weaver, Turing: Information theory, Bayesian inference

GREAT EXPECTATIONS

1954 Georgetown - IBM experiment Russian to English

Mi pyeryedayem mislyi posryedstvom ryechyi.
We transmit thoughts by means of speech.

- Translated 60 sentences.
- Claimed that MT would be solved within three or five years.
- Difference between limited and open domain.

DISAPPOINTMENT

1966 ALPAC report

- Concluded that MT was too expensive and ineffective
- Recommended that research focus on tools to help human translators

STATISTICAL RENAISSANCE

1993 Brown et al., ‘The mathematics of statistical machine translation’

The Fundamental Equation of Statistical Machine Translation

$$\hat{\mathrm{e}} = \underset{\mathrm{e}}{\operatorname{argmax}} \Pr(\mathrm{e})\Pr(\mathrm{f}|\mathrm{e})$$

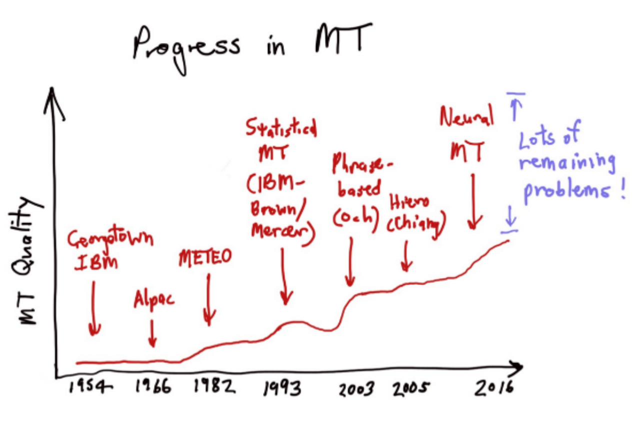
DATA

2000s Huge amounts of naturally occuring parallel data

GPUS

GPUs with LSTMs and other robust recurrent neural networks

PROGRESS SO FAR



(From Chris Manning’s slides)

A PLAN

DATA DRIVEN MACHINE TRANSLATION

- Specify a simple statistical model of translation

DATA DRIVEN MACHINE TRANSLATION

- Specify a simple statistical model of translation
- Learn the parameters of the model from data

DATA DRIVEN MACHINE TRANSLATION

- Specify a simple statistical model of translation
- Learn the parameters of the model from data
- Use linguistic analysis to inform and constrain the model

PARALLEL CORPORA

- Translated documents from governments, newspapers, etc.

PARALLEL CORPORA

- Translated documents from governments, newspapers, etc.
- What’s wrong with the data?

PARALLEL CORPORA

- Translated documents from governments, newspapers, etc.
- What’s wrong with the data?
  - It’s often noisy

PARALLEL CORPORA

- Translated documents from governments, newspapers, etc.
- What’s wrong with the data?
  - It’s often noisy
  - It’s in the wrong domains (mostly)

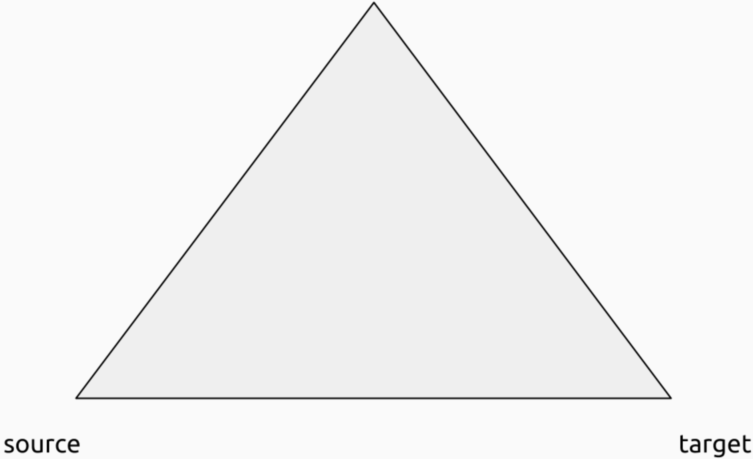
PARALLEL CORPORA

- Translated documents from governments, newspapers, etc.
- What's wrong with the data?
    - It's often noisy
    - It's in the wrong domains (mostly)
    - It's only partially observed

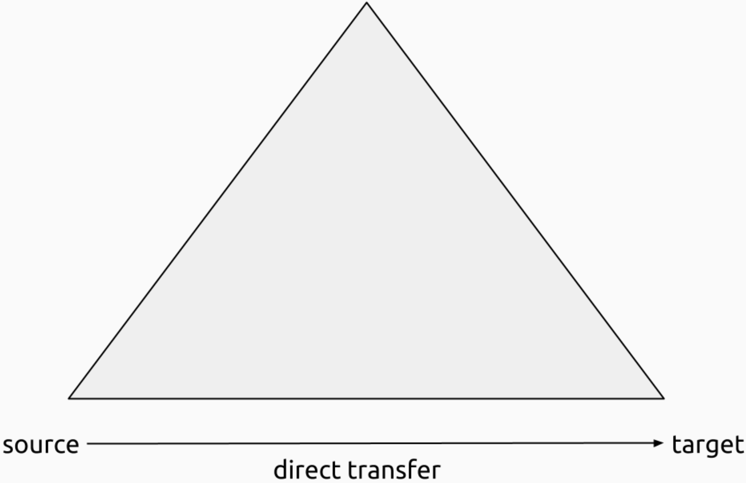
PARALLEL CORPORA

- Translated documents from governments, newspapers, etc.
- What’s wrong with the data?
    - It’s often noisy
    - It’s in the wrong domains (mostly)
    - It’s only partially observed
    - There’s not enough of it!

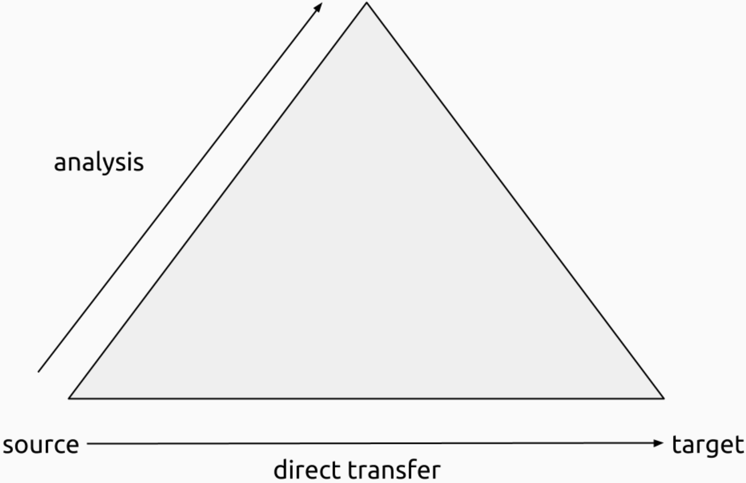
VAUQUOIS TRIANGLE



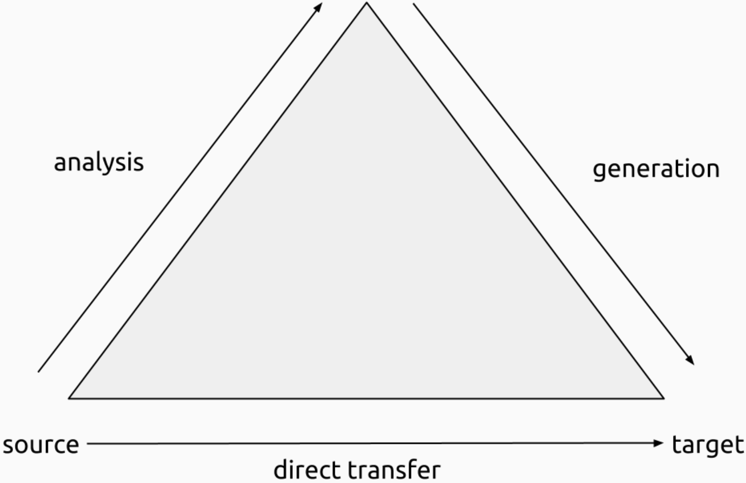
VAUQUOIS TRIANGLE



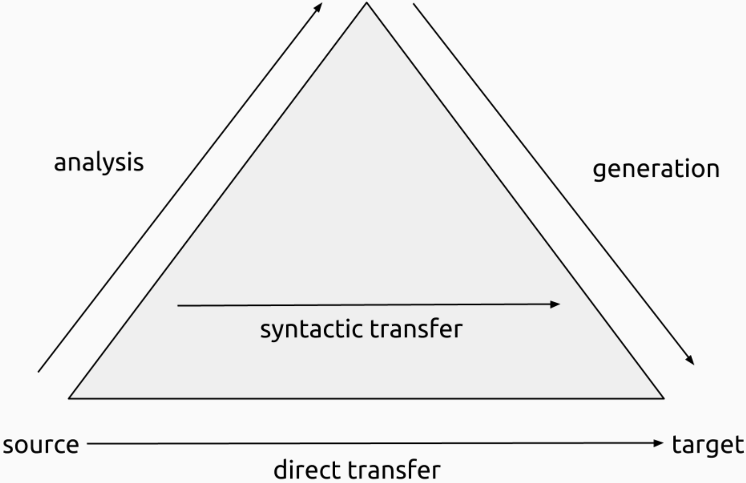
VAUQUOIS TRIANGLE



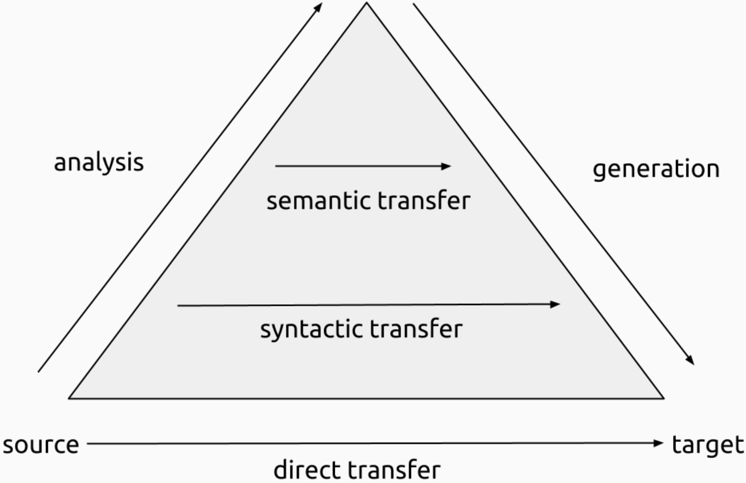
VAUQUOIS TRIANGLE



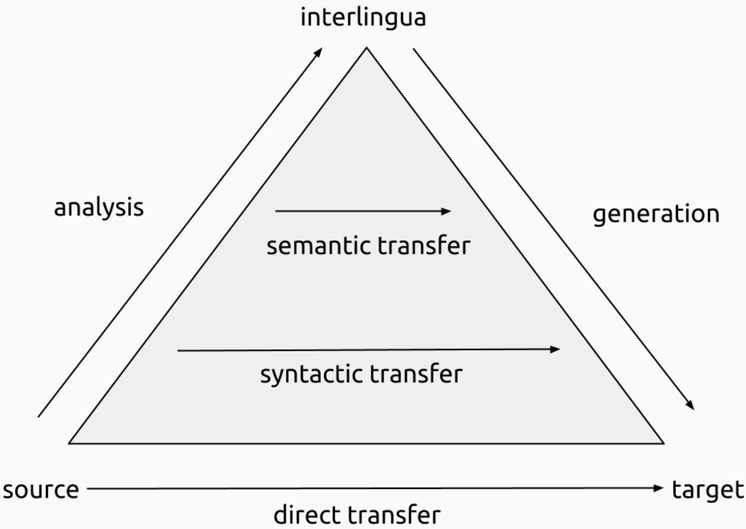
VAUQUOIS TRIANGLE



VAUQUOIS TRIANGLE



VAUQUOIS TRIANGLE



EVALUATION

HANDS-ON EVALUATION

Compare Bing, Google and Yandex Translate

- Work in pairs
- Compare sentences from random wikipedia articles (en<->ru)
- Add the source sentence, translations and judgements to this spreadsheet: `https://goo.gl/TcG5MZ`