Introduction to the Machine Translation Course

Aarne Ranta

European Masters, University of Malta, 18-22 March 2013

Course overview

Day 1: Introduction

Day 1,2: Rule-based translation: GF

Day 2,3: Statistical translation

Day 3: Hybrid systems

Day 4: Hands-on with the projects

Introduction overview

Some history of MT

Methods: rule-based, statistical, hybrid; interlingua, transfer

Evaluating MT

What is easy and what is difficult

Goals for this course

History

Early history

Turing: one of the things a machine could do

Shannon, Weaver: cryptography

- Russian is encoded English

optimism

The first critiques

Bar-Hillel (1960): *the pen is in the box*

The ALPAC report (1966): MT is low quality, useless, too expensive

Kay: MT must be interactive

Knowledge-based systems

Systran: transfer rules

Meteo: domain-specific (weather reports)

Rosetta: interlingual (Montague grammar)

VerbMobil: speech translation (unification grammar, Prolog)

The return of statistics

IBM: French to English trained at the Hansards corpus of Canadian Parliament

Google translate: on-line, 60 languages, based on the IBM ideas

Bing: Microsoft's on-line translator

Giza++ and Moses: open-source software for statistical MT

Pendulum swung too far?

Church (2011): there's no more low-hanging fruit

Hybrid systems: find the best combination of linguistics and statistics

Apertium: rule-based translation for closely related languages

GF: interlingual translation based on shared semantics

Methods

Rule-based

Word to word (dictionary lookup)

Rearrangement (of words)

Structure to structure (hierarchic phrases, not just words)

Use of grammars: morphology, syntax, semantics

Statistical

Noisy channel: French is distorted English

Word alignment: find corresponding words by looking at parallel texts

Language model: n-grams (sequences of *n* words)

Phrase-based: from words to multiwords (*for example*, *in spite of*)

Training: building the model from data

Decoding: applying the model at run time

Hybrid

Language = structures + distribution

Don't guess if you know

Factored systems: from words to lemma+analysis pairs

Tree-based systems: probabilistic grammars

Transfer vs. interlingua

Transfer: rules for each language pair

Interlingua: use an intermediate language

- a pivot language (English, Esperanto)
- a meaning representation (formal logic)

For n languages, interlingua needs $2n$ components, transfer needs $n(n\text{-}1)$

Sharing effort: perform operations on interlingua level

Linked Wordnets as interlingua: 80% of words in one-to-one correspondance

Evaluating MT

Manual evaluation

Quality criteria

- grammaticality
- fidelity (meaning preservation)
- fluency

Measure

- post-editing effort
- edit distance

Automatic evaluation

Gold standard: typically a separate part of the training material

Word error rate: how many words don't match with gold standard

BLEU: match words and n-grams (sequences of n words)

Evaluation as training: set parameters to maximize the BLEU score

What is easy and what is difficult

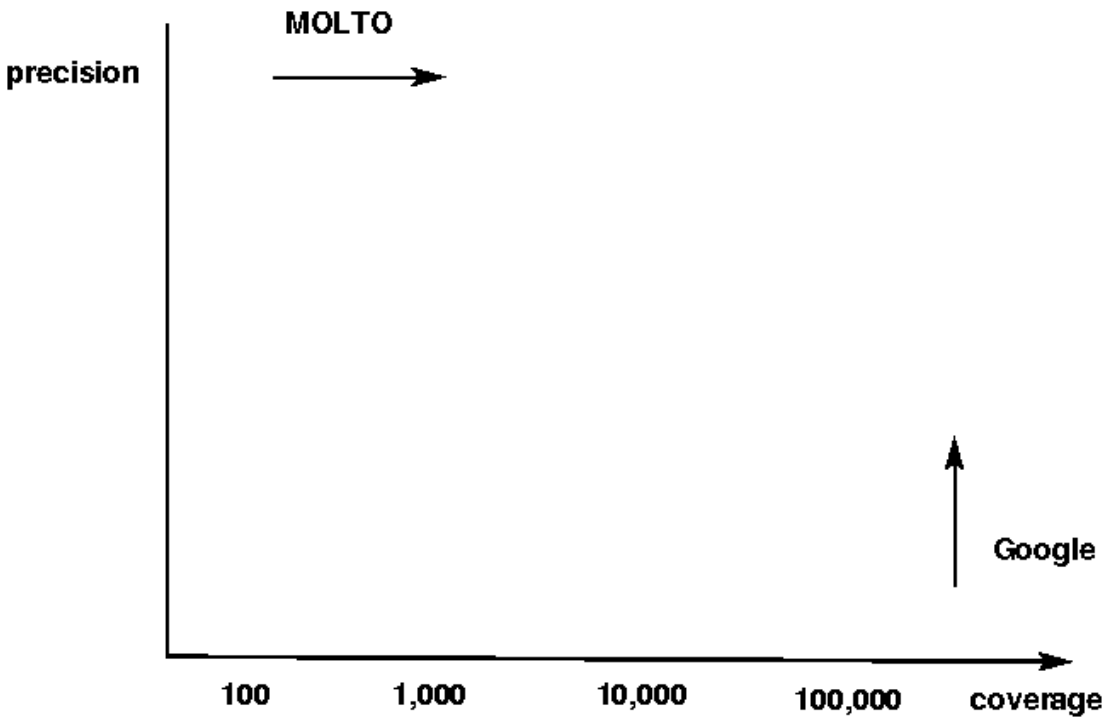
Depends on what you want

Coverage vs. **precision**

Browsing vs. **publication** (a.k.a **assimilation** vs. **dissemination**)

Bar-Hillel (1960): you cannot achieve both coverage and precision at the same time

Two systems and their ambitions



Coverage

Estimate of information needed: 100k words, 100M 2-grams, 10G 3-grams

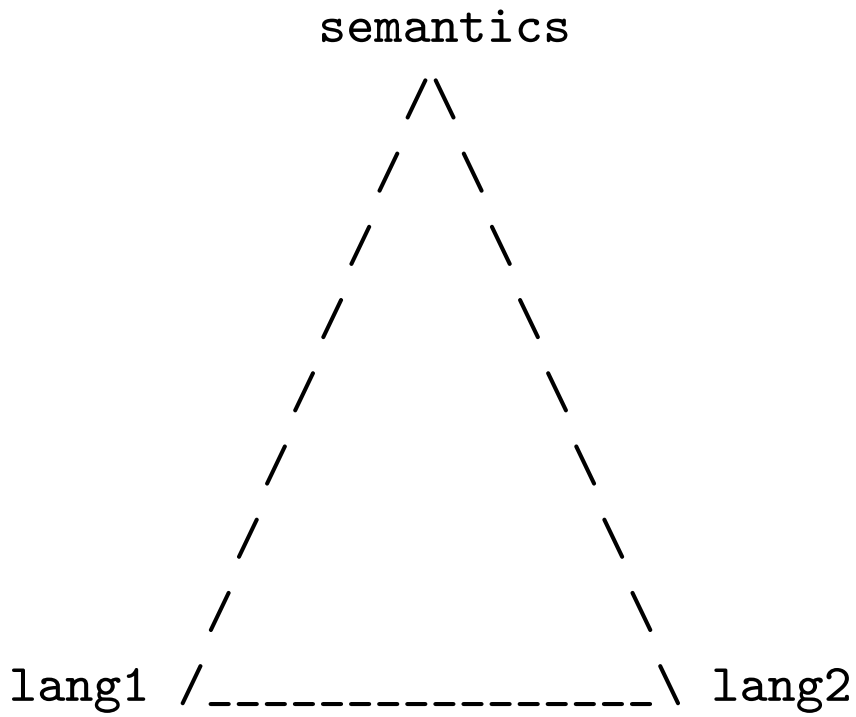
Morphological variation: 1000k word forms, 1000M 2-grams, 1000G 3-grams

Sparseness of data: hard to find all this

Smoothing: if you cannot find the 3-gram, combine two 2-grams

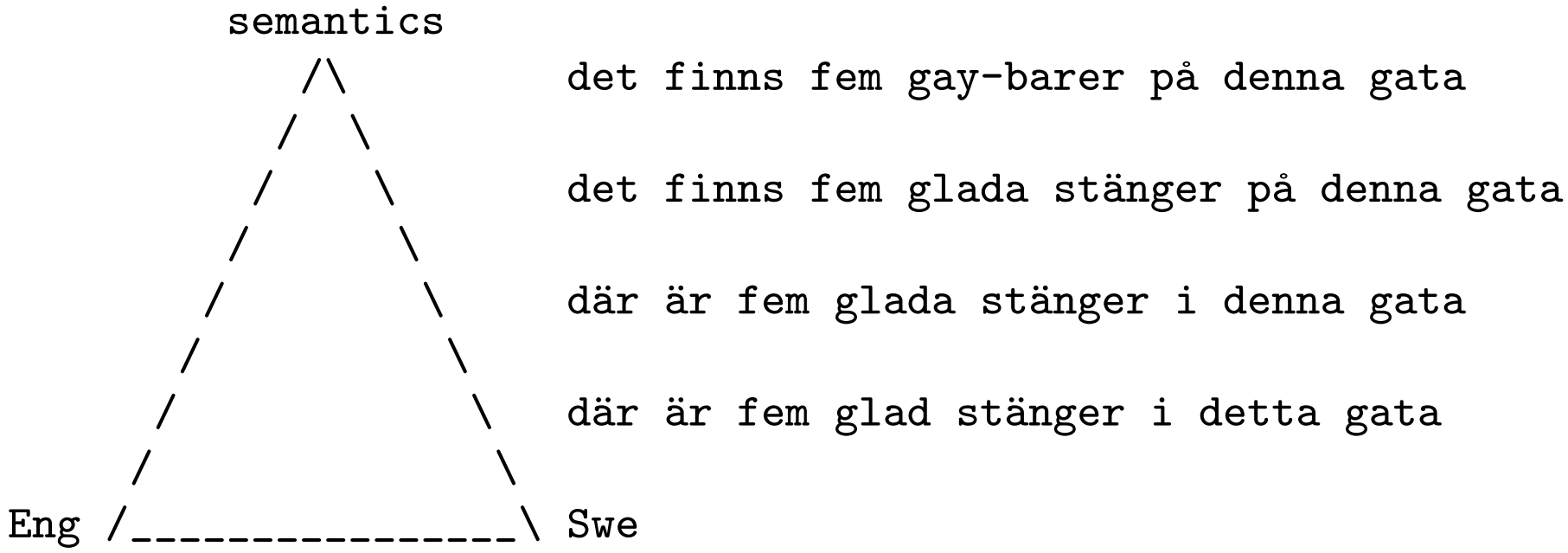
- *is here is* = *is here* + *here is*

The Vauquois triangle



Levels of analysis

there are five gay bars in this street



Long-distance dependencies

Agreement (French):

- *my father is intelligent - mon père est intelligent*
- *my mother is intelligent - ma mère est intelligent***e**
- *my mother is actually, regardless of what you say, very intelligent*

Discontinuous verbs (German):

- *er* **bringt** *dich* **um** - *he* **kills** *you*
- *er* **bringt** /deinen besten Freund// **um** - *he kills your best friend*

Reordering

The snow is white. If the snow is white, then the snow is white.

German: three orders,

Der Schnee ist weiss. Wenn der Schnee weiss ist, dann ist der Schnee weiss.

Disambiguation

I sent four **letters** *to the president*

I ate a pizza with shrimps
I ate a pizza with friends
I ate a pizza with chopsticks

Pros and cons of RBMT and SMT

Not just precision vs. coverage:

Grammatical correctness: RBMT

Meaning preservation: ?

Reordering: RBMT

Long distance; RBMT

Disambiguation: ?

Fluency: ?

Idioms, multiwords: SMT

Low-resourced languages: RBMT?

Effort needed: SMT

Predictability: RBMT

Programmability: RBMT

Ideal languages for SMT

Morphologically simple

Rigid word order

Lots of data

English! And Swedish, Dutch, French,...

Ideal languages for RBMT?

Morphologically simple

Free/varying word order

Lack of digital data

Notoriously bad for SMT: Finnish, Japanese,...

An ideal hybrid system?

Taking all pros and cons into account

Not easy

The goals of this course

Build systems using existing tools

- GF
- Giza + Moses

Understand what it easy and what is difficult

- estimate the effort for a GF project
- know how to use Google translate - and how to cheat it