## MT History

Machine translation has a long history:

- Warren Weaver (1947) suggested using new fangled Computers to translate between human languages.
- Most research in Computational Linguistics until 1966 was in MT.
- The ALPAC Report (1966):
  - Argued that research in Machine Translation would never produce high quality translations.
  - Resulted in US funding for MT being cancelled.
- Research in MT continued in the 1980s with rule-based systems:
  - Systran is a good example.
  - Requires a lot of manual crafting.



## Statistical Machine Translation

From the BBC Arabic Page Sunday $26^{th} November$:

> Olmert said that he is optimistic that the cease-fire will help in the resumption of talks on the stalled peace process between Israel and the Palestinian Authority and help significantly to the release of the Israeli soldier "Gilead Shalit," which Palestinian gunmen detained several months ago.

http://www.google.co.uk/language_tools

## SMT

Current Statistical Machine Translation builds upon the IBM models:

- A basic insight is that whole sequences of words can be translated (rather than translating word-by-word).
  - Whole sequences are called *phrases*:

    *The cat*
- Translating using phrases has advantages:
  - Non-compositional units can be translated.
  - There is no need to translate these phrases themselves.

The resulting models are called *Phrase-based*.



## MT History

In the early 1990s, IBM developed a statistical approach (Candide):

- Parallel corpora was used to automatically determine how individual words are translated.
- Translation was a matter of decoding.
- The system made minimal use of linguistics.

*Every time I fired a linguist my performance goes up*

## SMT

The major components of SMT:

- A *language model*:
  - A LM captures how fluent a translation is.
  - LMs are trained using monolingual data.



## SMT

The major components of SMT:

- A *phrase-table*:
  - This captures all the possible translations of each source phrase.
  - Translations are also phrases.
- Phrase tables are created heuristically using the word-based IBM models.

## SMT

In more detail:

- The probability of a target phrase is:

$$P(e \mid f) = \frac{\text{freq(e,f)}}{\text{freq(f)}}$$

- Here, we gather counts from a word-aligned parallel corpus.
- $e$ is a target phrase and $f$ is a source phrase.
- (It can be useful to model $P(f \mid e)$)
- This models *adequacy*.



## SMT

The major components of SMT:

- A *decoder* actually searches for the best translation.
  - It uses the phrase tables and language models.
  - It also *reorders* the target sentence.
- Typically, the search space is so large that a *beam search* is necessary.

## SMT

In more detail:

- The probability of a particular target translation:

$$P(e_1, e_2, \ldots, e_n) = P(e_1) \cdot P(e_2 \mid e_1) \cdot P(e_3 \mid e_1, e_2) \ldots$$

- This is a language model and counts can be gathered from monolingual data.
- This models *fluency*.



## SMT

Using the Noisy Channel model:

$$P(E \mid F) = P(E) \prod_e P(f \mid e)$$

A decoder searches for the best target translation $E$, given a source sentence $F$.

## SMT

A log-linear model:

$$E^* = \text{argmax}_E \sum_i f_i(E,F)\lambda_i$$

- Now, all components are feature functions.
- The weights balance the contribution each part makes.
- The weights are set to maximise translation performance on a development set.



## SMT

A log-linear model:

- It can also be useful to add other sources of information into the mix:
  - How good a candidate translation is some phrase?
  - Are we producing translations that are too long or short?
- Also, we may wish to differentially weight each part:
  - The language model is likely to be much better estimated than the translation model.
  - Not all components are equally important as each other.
- It can be useful to weight components such that performance on some metric is optimised.

## SMT

Open questions:

- A major question is whether *syntax* can improve performance.
  - Arguably, translation is locally good, but globally poor:

    For his part, Israeli Defense Minister Amir Beriz that the Israeli army would continue its operations in the Gaza Strip, and Palestinian militants continued to fire rockets towards Israel.
  - Only until very recently have syntax-based translation systems been shown to be on a par with phrase-based systems.

*Every time I fired a Statistician I get a warm fuzzy feeling*



## SMT

Example features:

$$f_1 = \left\{ \; P(\text{e} \mid \text{f}) \right.$$

$$f_2 = \left\{ \; P(\text{f} \mid \text{e}) \right.$$

$$f_3 = \left\{ \; P(e_i, e_{i+1}, e_{i+2}) \right.$$

## SMT

Open questions:

- SMT models are mainly generative:
  - Model weights really act as scaling factors.
  - Building large-scale discriminative models –with millions of features– is a hard problems.



## SMT

Open questions:

- Evaluation is tricky:
  - The current dominant approach uses string similarity against a set of reference translations (Bleu).
  - Bleu has been shown to be biased towards SMT systems (as opposed to rule-based approaches such as Systran).
  - Human evaluation is the best, but it is expensive.

## SMT

Open questions:

- Most SMT research has focused on Chinese-English and Arabic-English.
  - Languages with much less training material are not so heavily researched.
- Can we build models which have better generalisation capabilities?



## SMT

Open questions:

- Performance is reaching a plateau:
  - Better performance comes from using more training material.
  - Language models can be improved using trillions of words.
  - How can we improve our phrase models knowing that the amount of parallel material is much less?

## Comments

SMT is currently a hot, well-funded area:

- Many problems to work-on.
- A good chance to get a job doing this later.

The Edinburgh SMT group will be setting a number of MSc projects this summer.