SMT – TIDES – and all that

Aus der Vogel-Perspektive
A Bird’s View (human translation)

Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
Machine Translation Approaches

- Interlingua-based
- Transfer-based
- Direct
  - Example-based
  - Statistical
Statistical versus Grammar-Based

- Often statistical and grammar-based MT are seen as opposing approaches – wrong !!!

- Dichotomies are:
  - Use probabilities – everything is equally likely (in between: heuristics)
  - Rich (deep) structure – no or only flat structure

- Both dimensions are more or less continuous

- Examples
  - EBMT: flat structure and heuristics
  - SMT: flat structure and probabilities
  - XFER: deep(er) structure and heuristics

- Goal: structurally rich probabilistic models
**Statistical Approach**

- **Using statistical models**
  - Create many alternatives (hypotheses)
  - Give a score to each hypothesis
  - Select the best -> search

- **Advantages**
  - Avoid hard decisions, avoid early decisions
  - Sometimes, optimality can be guaranteed
  - Speed can be traded with quality, no all-or-nothing
  - It works better! (in many applications)

- **Disadvantages**
  - Difficulties in handling structurally rich models, mathematically and computationally (but that’s also true for non-statistical systems)
  - Need data to train the model parameters
Statistical Machine Translation

Based on Bayes’ Decision Rule:

$$\hat{e} = \arg\max\{ p(e \mid f) \} = \arg\max\{ p(e) \cdot p(f \mid e) \}$$
Tasks in SMT

- Modelling
  build statistical models which capture characteristic features of translation equivalences and of the target language

- Training
  train translation model on bilingual corpus, train language model on monolingual corpus

- Decoding
  find best translation for new sentences according to models
Alignment Example

- Translation models based on concept of alignment
- Most general: each source word aligns (partially, with some probability) to each target word
- Additional restrictions to make it mathematical and computationally tractable
Translation Models

• The heritage: IBM
  • IBM1 – lexical probabilities only
  • IBM2 – lexicon plus absolut position
  • IBM3 – plus fertilities
  • IBM4 – inverted relative position alignment
  • IBM5 – non-deficient version of model 4

• In the same mood:
  • HMM – lexicon plus relative position
  • BiBr – Bilingual Bracketing, lexical probabilities plus reordering via parallel segmentation
  • Syntax-based – align parse trees
Training

- Need bilingual corpora
  - Usually, the more the better
  - But needs to be appropriate – domain specific - and clean
  - No need for manual annotation

- Training of word alignment models
  - Iterative training: EM algorithm
  - For HMM: Forward-Backward
  - For BiBr: Inside-Outside
  - Often maximum approximation: Viterbi alignment

- GIZA toolkit
  - Partly developed at JHU workshop
  - Chief programmer: Franz Josef Och
How does it work?

- First iteration: start with uniform probability distribution

| Bilingual Corpus: | Word Pairs: | Probabilities $p(s|t)$: |
|-------------------|-------------|-------------------------|
| A B C # R S T     | A - R : 2   | A - R : 2/7             |
| E B F G # S U V   | A - S : 2   | A - S : 2/11            |
| A D B E # R V S   | A - T : 1   | A - T : 1/3             |
|                   | B - R : 1   | B - R : 1/2             |
|                   | B - S : 3   | B - S : 3/11            |

- Next iteration: multiply counts by probabilities always renormalize
Phrase Translation

- **Why?**
  - To capture context
  - Local word reordering

- **How?**
  - Typically: Train word alignment model and extract phrase-to-phrase translations from Viterbi path
  - But also: Integrated segmentation and alignment
  - Also: rule-base segmentation

- **Notes:**
  - Often better results when training target to source for extraction of phrase translations due to asymmetry of alignment models
  - Phrases are not fully integrated into alignment model, they are extracted only after training is completed
Language Model

- Standard n-gram model:
  \[ p(w_1 \ldots w_n) = \prod_i p(w_i \mid w_1 \ldots w_{i-1}) = \prod_i p(w_i \mid w_{i-2} w_{i-1}) \text{ trigram} = \prod_i p(w_i \mid w_{i-1}) \text{ bigram} \]

- Many events not seen -> smoothing required

- Also class-based LMs and syntactic LMs, interpolated with word-based LM

- Use of available toolkits: CMU LM toolkit, SRI LM toolkit
Search for the best Translation

- Given new source sentence
- Brute force search
  - Translation model generates many translations
  - Each translation has a score, including the language model score
  - Pick the one with the highest score
- Result
  - Best translation according to model
  - Not necessarily the best translation according to evaluation metric
  - Not necessarily the best translation according to human judgment
- Realistic search
  - ‘Grow’ many translations in parallel
  - Throw away low scoring candidates (pruning)
  - Search errors: found translation is not the best according to models
MT Evaluation

- Human evaluation – all along
  - Fluency, adequacy, overall score, etc.
  - Problems: inter-evaluator agreement, reproducibility, cost

- Automatic scoring
  - Use one or several reference translation to compare againsts
  - Define a distance measure, then: the closer, the better

- Different scoring metrics proposed and used
  - Position independent error rate (how many words are correct)
  - Word error rate (are the all in the correct order)
  - Blue n-gram: how many n-grams match
  - NIST n-gram: how many n-grams match, how informative are they
  - Precision – Recall

- MT Evaluation – hot topic, more competition in metric development than in MT development
DARPA funded NLP project:

T – Translingual  (Translation undercover ;-)  
I – Information  
D – Detection  
E – Extraction  
S – Summarization

- Large number of research groups (universities and companies)

Program Objective

- Develop advanced language processing technology to enable English speakers to find and interpret critical information in multiple languages without requiring knowledge of those languages.
Program Strategy

- **Research**
  Conduct research to develop effective algorithms for detection, extraction, summarization, and translation -- where the *source data may be large volumes* of naturally occurring speech or text in multiple languages.

- **Evaluation**
  Measure accuracy in *rigorous, objective evaluations*. Outside groups are invited to participate in the annual Information Retrieval, Topic Detection and Tracking, Automatic Content Extraction, and Machine Translation evaluations run by NIST.

- **Application**
  Integrate core capabilities to form effective text and audio processing (TAP) systems. Experiment with those systems on *real data with real users*, then refine and iterate.
MT in TIDES

- Evaluations every year
  - Chinese large data track: > 100m words of bilingual corpus
  - Chinese small data track: 100k words bilingual corpus, 10k dictionary
  - Arabic large data track: 80m words bilingual corpus
  - Open data track: use whatever you can find before data collection deadline – but no significant improvement over large data track results

- Many strong teams
  - TIDES funded plus external groups
  - Friendly competition: you tell me your trick – I tell you my trick

- Exciting improvements over last two years
  - Automatic metrics over-score machine translations or under-score human translations
**Surprise Language Evaluation**

- Do learning approaches allow to build useful NLP system for new language within weeks?

- **Dry run exercise: Cebuano**
  - Only data collection
  - Most data essentially found within days
  - Very inhomogeneous corpus resulted: Bible to party propaganda

- **Actual evaluation: Hindi**
  - Enormous problems with different encodings, many proprietary
  - Amount of data > 2 million words bilingual
  - Several dictionaries
  - MT systems, but also NE tagging, cross-lingual IR, etc built within 4 weeks
  - Nobody liked it: only dealing with encoding, no new NLP research
The Future

- Continuous evaluations: Arabic and Chinese and perhaps new surprises
- Possible other genres, not only news
- Constant improvements
  - In evaluation approaches ;-)  
  - But also in translation !
- Similar comparative evaluations are underway and will follow in other projects, also for speech-to-speech translation