technology
from seed

The INESC-ID Phrase-based Statistical Translation System

Diamantino Caseiro
Isabel Trancoso

Outline

- Statistical framework
- System Architecture
- Weighted Finite-State Transducers
- Decoding
- System Training
- Models
- Results
- Conclusions
- Future Work

Statistical Framework

technology
from seed

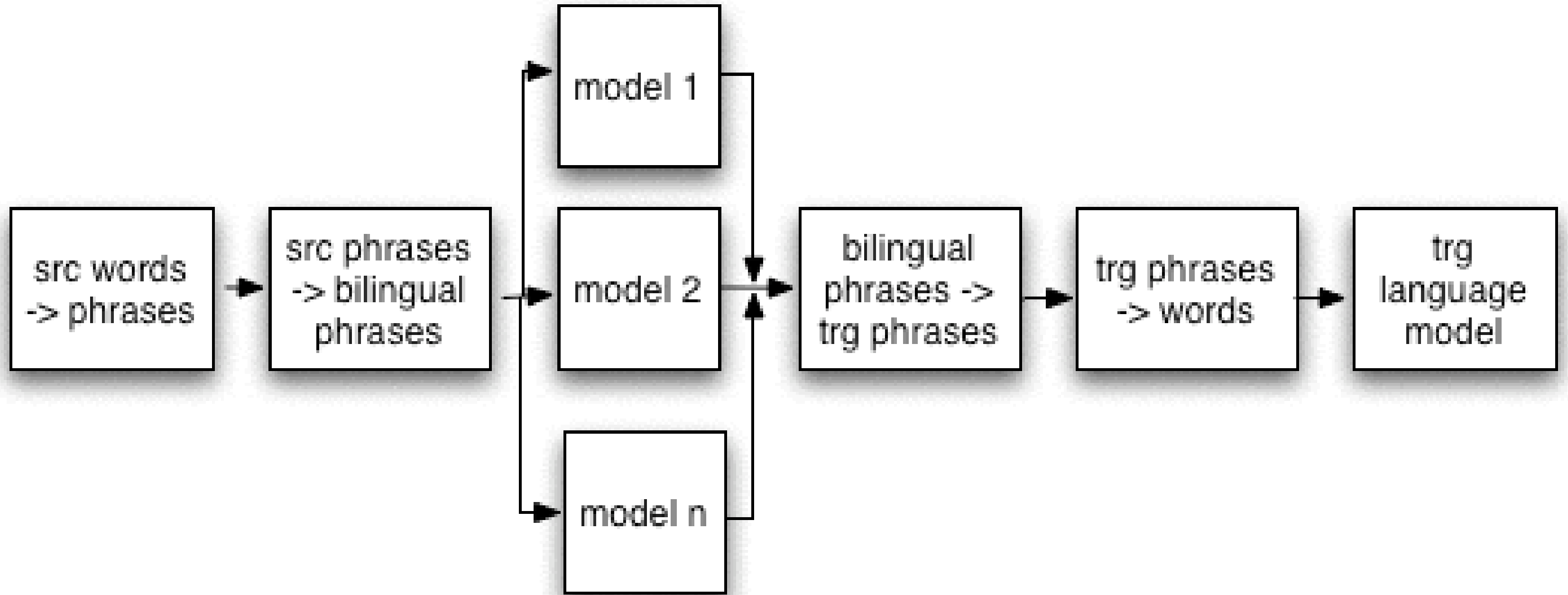
- Log-linear Model

$$p(e_1^I \mid f_1^J) = \frac{1}{Z(f_1^J)} \exp\left(\sum_{n=1}^{N} \lambda_n h_n(e_1^I \mid f_1^J)\right)$$

- Allows for easy integration of multiple knowledge sources
- Parameters can be tuned to the desired objective function

System Architecture

- Based on Weigthed Finite-State Transducers (WFST)



System Architecture

technology
from seed

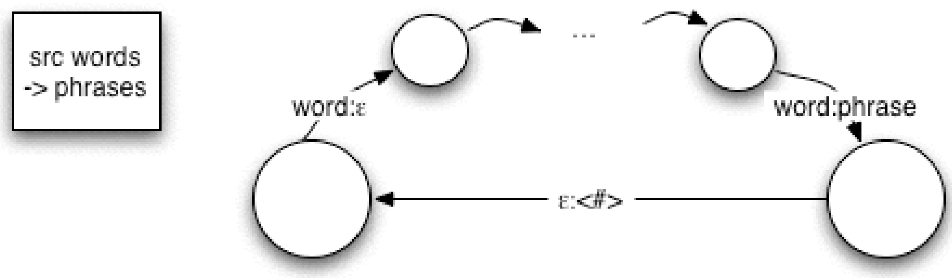
inesc id lisboa

- Each module is represented by a WFST:
  - S input sentence
  - L_s maps source words to source phrases
  - M_s maps source phrases to bilingual phrases
  - λ_i is a log linear scalar parameter
  - H_i is a translation model (bilingual phrase log probability table) represented by an automaton
  - M^i_t maps bilingual phrases to target phrases
  - L^i_t maps target phrases to target words
  - λ_g is the language model log linear scale
  - G_t is a target language model

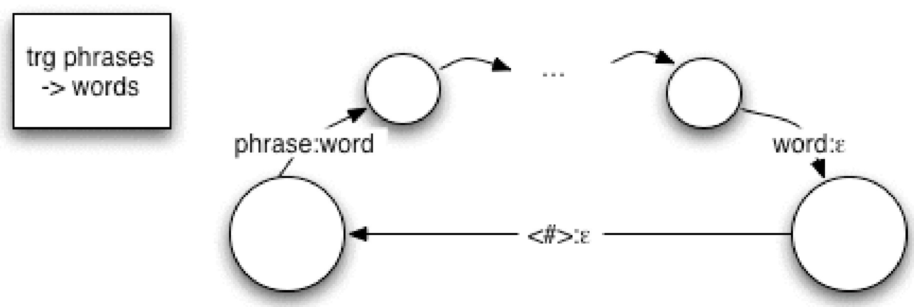
- $\text{viterbi}[S \circ L_s \circ M_s \circ (\lambda_1 H_1 \cap \ldots \cap \lambda_n H_n) \circ M^i_t \circ L^i_t \circ \lambda_g G_t]$

Transducers

- L_s src words-> src phrases
  - Tree structure

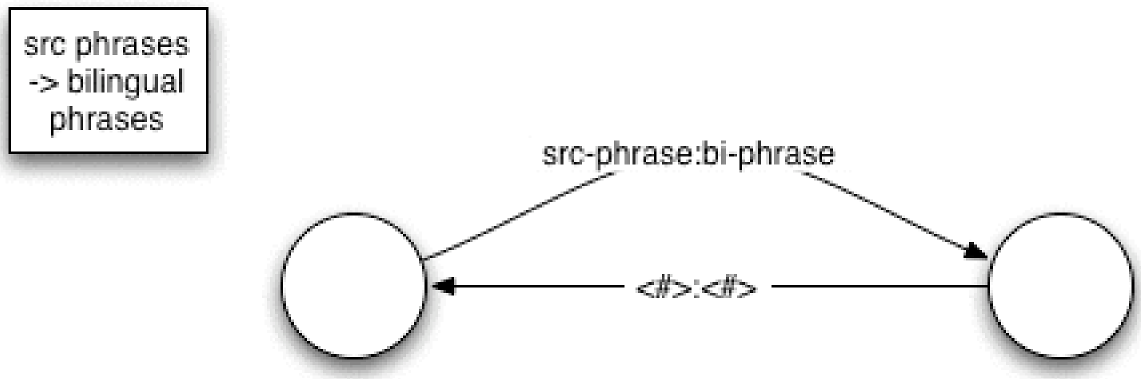


- L^i_t trg phrases-> trg words
  - Reverse tree structure
  - <#> “end-of-phrase” symbol is used for composition efficiency

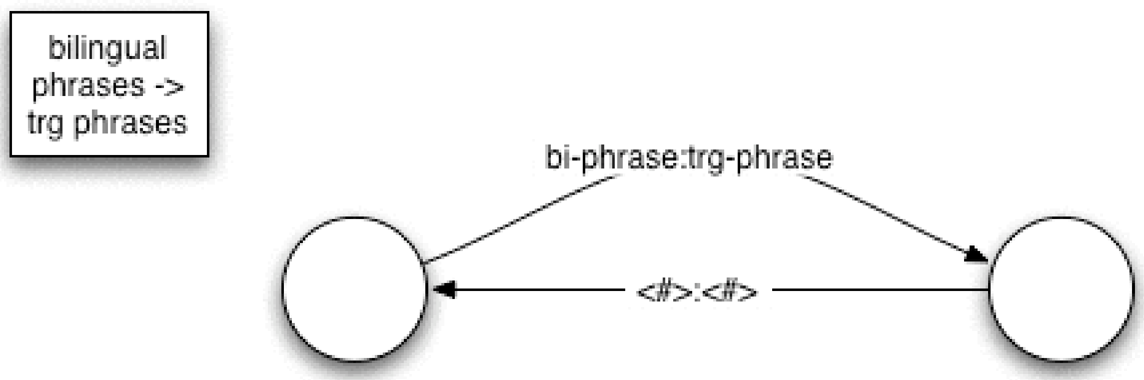


Transducers

- M_s src phrase -> bi phrase

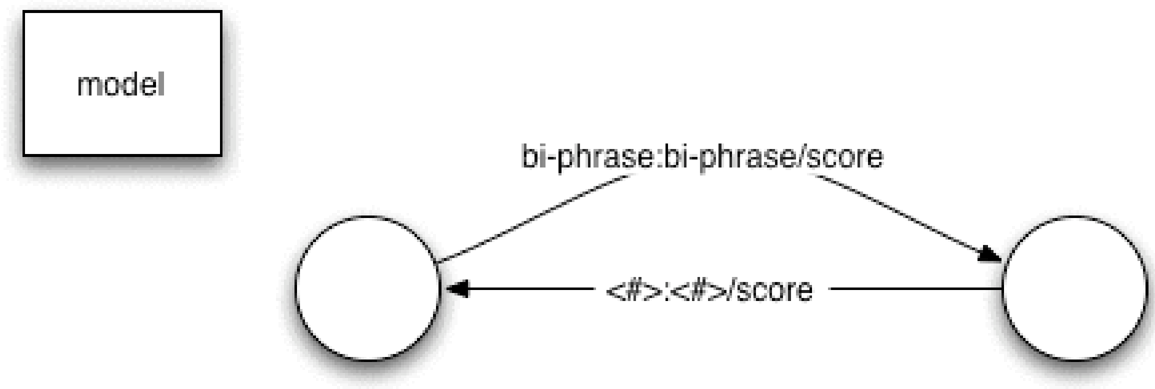


- M^i_t bi phrase -> trg phrase



Transducers

- H_i translation models



- Viterbi, source-word synchronous search:
  - "On-the-Fly" transducer composition
  - Caching
  - Beam pruning
  - Histogram pruning
  - Lattice generation
    - pruning + optimization
  - Transducers can be integrated offline in a static network, however:
    - "On-the-Fly" transducer composition is used during development to allow tuning of parameters without recreating the network

Decoding

technology from seed

- Monotonic search
  - No reordering is currently being applied
- Implemented using INESC-ID in-house finite-state transduter toolkit (fstk)
- Around 3 sentences per second on a Pentium 4 3.2Ghz with 2GB Ram (using file mapped transducers)

System Training

- Sentence Selection
  - Less than 55 words in either language
  - Less than 1.3 fertility
- Alignments:
  - GIZA++ alignments
    - (5x IBM1, 9x HMM, 5x IBM3, 5x IBM4)
  - Only Spanish->English alignments were used
  - Heuristics:
    - Target words aligned to NULL are merged with the previous output word (except the first, that is merged with the second word)
    - Source word with no alignment are merged with the next source word
    - Merged words are allways kept in the same phrase

- Phrase Selection
  - Up to 5 source words
    - (merged words are considered as a single word)
  - Must occur at least twice
  - Single word alignments can occur only once
  - Must be auto contained
    - Alignment with a word outside the phrase is not allowed
  - 6 Million phrases selected

Models

- Phrase Model

$$p(f \mid e) = \frac{N(f,e)}{N(e)}$$

- IBM1 Lexical Model

$$p(t \mid s) = \frac{1}{(I+1)^J} \prod_{J=1}^{J} \sum_{i=0}^{I} p_{ibm1}(t_i \mid s_j)$$

Models

- Target-word insertion penalty
- Phrase insertion penalty
- Language model
  - 4-gram (3.7M 4-grams + 9.3M 3-grams + 2.1M 2-grams)
  - SRILM (modified Knesser-Ney)
- Parameters were chosen to minimize Bleu
  - Downhill simplex algorithm (Math:Amoeba perl package)

Results

WER	PER	Bleu	Nist
38.07	28.93	48.81	9.99

Conclusions

- WFST implementation of phrase-based SMT
- “On-the-Fly” composition allows a modular architecture
- Prelimitary translation results are encoraging

Future Work

technology
from seed

inesc id lisboa

- Perform more experiments:
  - Evaluate the contribution of each individual model
- Use bidirectional alignments
- Use morphological information
  - Lematizer and POS-taggers can be easily integrated in the transducer cascade (we built n-gram based taggers with less that 2% error, however its use did not improve the system)
- Train and evaluate using speech recognition output
- Explore tight integration with a WFST based speech recognizer
- Phrase reordering


technology
from seed
L² F - Spoken Language Systems Laboratory