INF5820/INF9820

LANGUAGE TECHNOLOGICAL APPLICATIONS

Jan Tore Lønning, Lecture 9, 17 Oct. 2014

jtl@ifi.uio.no

Today

- Generative vs Discriminative
- Hybrid translation: Rule based + discriminative training
  - Treebanks and parse ranking
  - Generation ranking
  - Ranking end-to-end
- Reranking in statistical MT
- A glimpse beyond

Generaitve modelling

- Make a model of how the data are produced (generated)
- Split it up in smaller steps
- Assign probabilities:
  - To the steps
  - Calculate them together to a probability for the data
- Use this to select the (n-)best candidates of how the data are generated
- Examples:
  - Probabilistic context-free grammars
  - Statistical Machine Translation

Discriminative training

- Consider candidate solutions
  - (coming from somewhere)
- Have some way to evaluate them
  - Some score, or ranking
  - Or supervised training material
- Choose features
- Use machine learning to select the best from the features
- Examples:
  - Malt parser (parser without grammar)
  - Parse ranking
  - Ranking of rule-based MT
  - Reranking in SMT

Today

- Generative vs Discriminative
- Hybrid translation: Rule based + discriminative training
  - Treebanks and parse ranking
  - Generation ranking
  - Ranking end-to-end
- Reranking in statistical MT
- A glimpse beyond

Parse ranking

- First build a parse bank
  - Demo on
  - http://clarino.uib.no/iness/page
  - (http://erg.delph-in.net/logon)
- Then use this for building a discriminator to select/rank between candidates
- Choices:
  - Features
  - Learning algorithm

Compare to prob. grammars

Prob. Grammar (PCFG)

- Generative model
- Construct parses
- Rank the parses
- Use grammatical/tree features

Parse ranker

- Discriminative model
- Select between candidate parses constructed elsewhere
- Large freedom in use of features

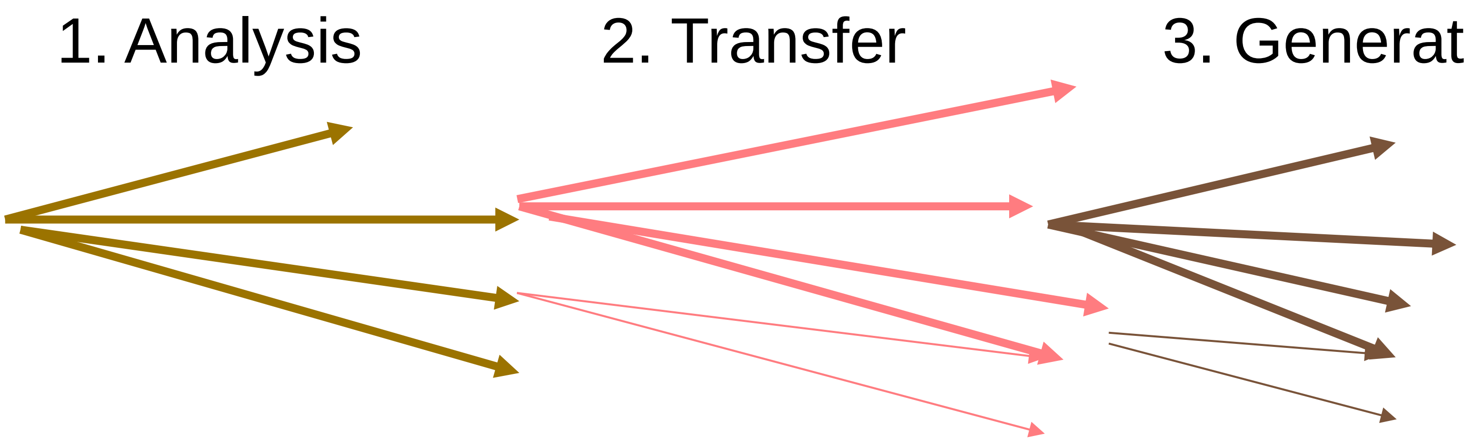
Generation ranker

- Roughly 30 realizations per MRS
- First attempt:
  - N-gram language model

model	exact match	five-best	WA
BNC LM	53.24	78.81	0.882
Log-Linear	72.28	84.59	0.927

- Better:
  - Inspired by parse ranking
  - Developed on the basis of a parse bank
  - Extract features
  - Max-ent learning
  - Better results!

Ambiguity

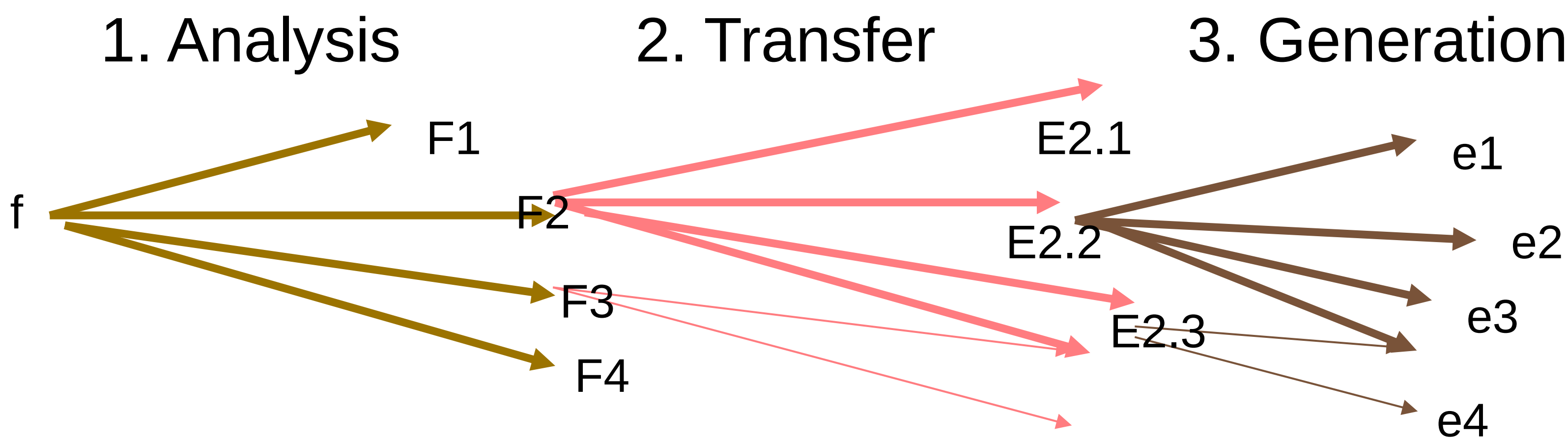


- Stochastic models score the alternative outcomes of each component: Parsing, Transfer, Generation
- The per-component scores are calculated together and the final outcomes are ranked.
- Component models are trained on corpora and treebanks.

Transfer

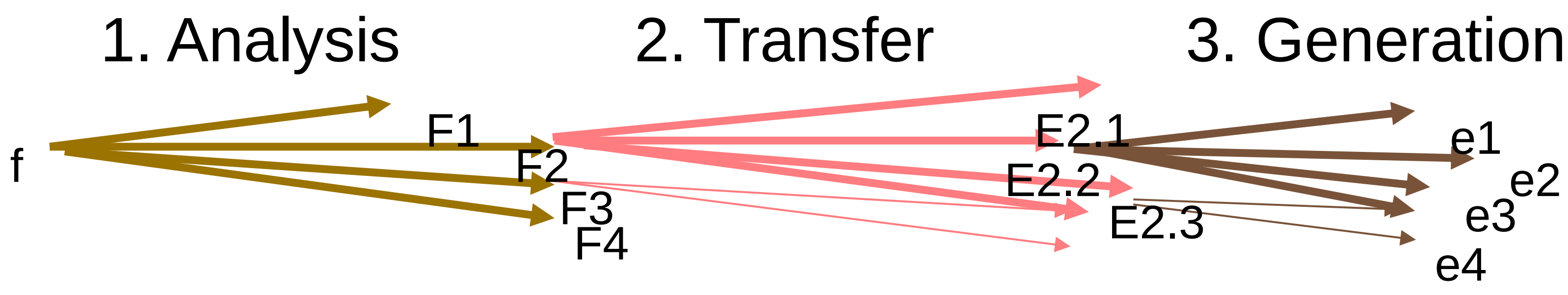
- Should have been **conditional** probabilities:
  - The probability of an English MRS given a Norwegian MRS:
- Only included absolute probabilities:
  - The probability of an English MRS

Putting the 3 together



- Alternatives
  1. First $\arg\max_i (F_i \mid f)$, say F_2, then $\arg\max_j (E_j \mid F_2)$ etc
  2. The most likely path $\arg\max_{i,j,k} P(e_k \mid E_j)(E_j \mid F_i)(F_i \mid f)$
  3. The most likely translation $\arg\max_e \sum_{F_i} \sum_{E_j} P(e_k \mid E_j)(E_j \mid F_i)(F_i \mid f)$

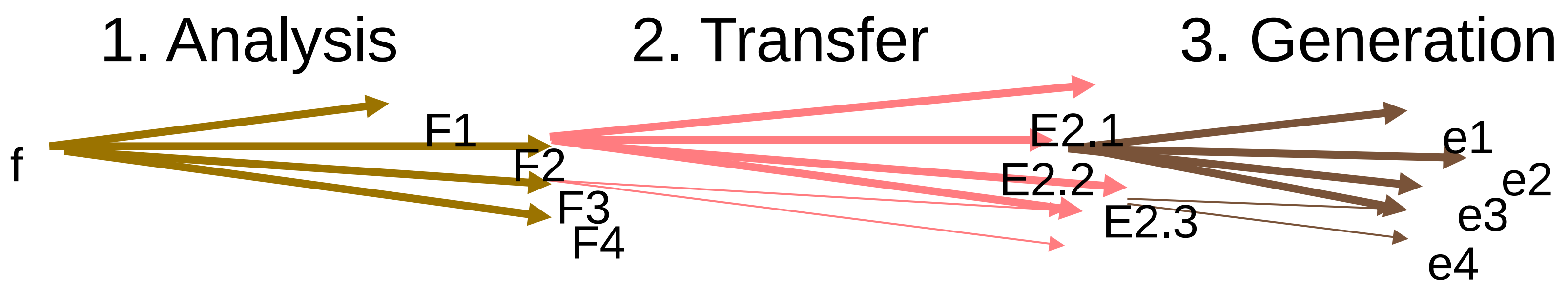
Putting the 3 together



1. First $\arg\max_i (F_i \mid f)$, say F_2, then $\arg\max_j (E_j \mid F_2)$ etc

- Theoretically sound:
  - The best parse is in principal independent of the translation, etc.

Putting the 3 together

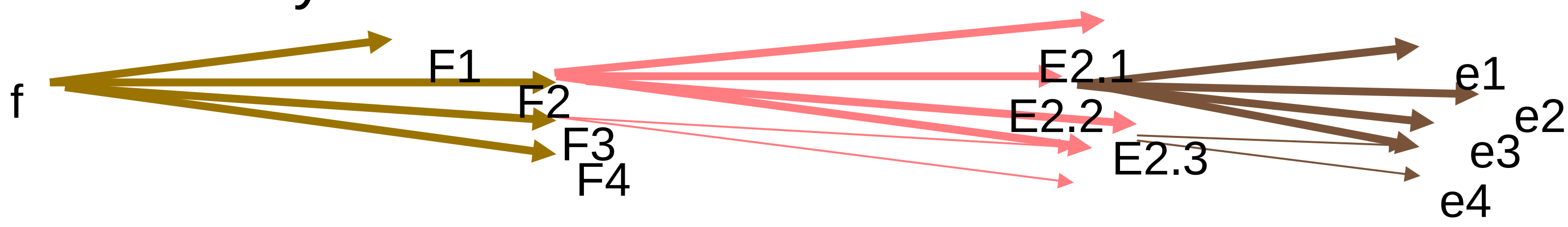


2. The most likely path $\arg\max_{i,j,k} P(e_k \mid E_j)(E_j \mid F_i)(F_i \mid f)$

- Might yield better results:
  - When we see that the translation is unlikely, we may detect mistakes earlier in the process

Putting the 3 together

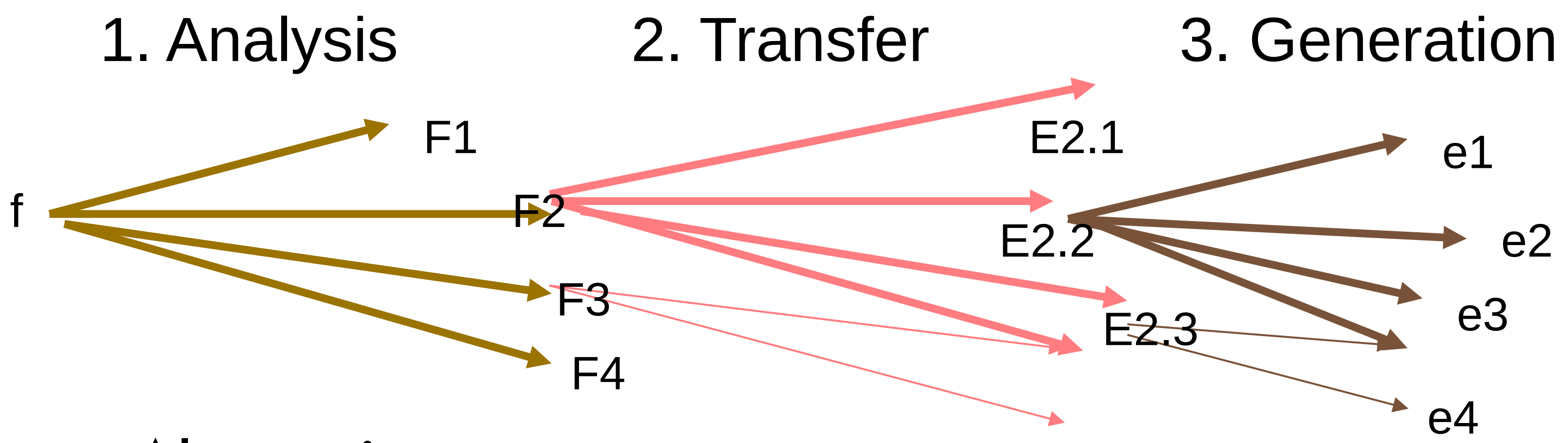
1. Analysis 2. Transfer 3. Generation



3. The most likely translation $\arg\max_{e} \sum_{F_i} \sum_{E_j} P(e_k \mid E_j)(E_j \mid F_i)(F_i \mid f)$

- Might yield better results:
  - Ambiguities in source language may be the same in target language, e.g. PP-attachement
    - Jeg så mannen i parken med kikkerten
    - I saw the man in the park with the binoculars
    - The same 5 way ambiguity in Norw. and English

Putting the 3 together



- Alternatives
  1. First $\arg\max_{i} (F_i \mid f)$, say F_2, then $\arg\max_{j} (E_j \mid F_2)$ etc
  2. The most likely path $\arg\max_{i,j,k} P(e_k \mid E_j)(E_j \mid F_i)(F_i \mid f)$
  3. The most likely translation $\arg\max_{e} \sum_{F_i} \sum_{E_j} P(e_k \mid E_j)(E_j \mid F_i)(F_i \mid f)$

End-to-end reranking

- Why?
  - Possibly correct the individual modules
  - More information
  - Similar to model 3 on last slide
- Features:
  - The 3 modules
  - Lexical trans. probabilities
  - Word order etc.

Results

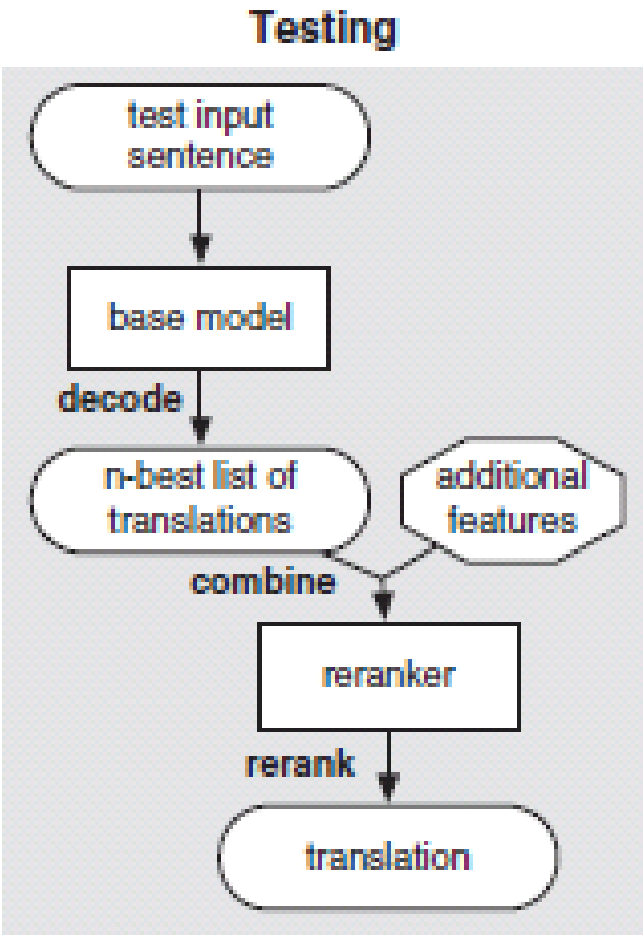
set	#	chance	first	LL	top	judge
JH_d	1391	34.18	40.95	44.10	49.89	–
JH_t	115	30.84	35.67	38.92	45.74	46.32

Table 4: BLEU scores for various re-ranking configurations, computed over only those cases actually translated by LOGON (second column). For all configurations, BLEU results on the training corpus are higher by about four points.

Today

- Generative vs Discriminative
- Hybrid translation: Rule based + discriminative training
  - Treebanks and parse ranking
  - Generation ranking
  - Ranking end-to-end
- Reranking in statistical MT
- A glimpse beyond

Reranking model for SMT



- Discriminative model
- Take as input an n-best list from a translation system

Reranking vs Tuning

- What is the difference between
  - Tuning and
  - Reranking?

Today

- Generative vs Discriminative
- Hybrid translation: Rule based + discriminative training
  - Treebanks and parse ranking
  - Generation ranking
  - Ranking end-to-end
- Reranking in statistical MT
- A glimpse beyond

A glimpse beyond: Minimum Bayes Risk

$$\mathbf{e}_{\text{best}}^{\text{MAP}} = \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}, \mathbf{a}|\mathbf{f}) \tag{9.36}$$

$$\mathbf{e}_{\text{best}}^{\text{SUM}} = \text{argmax}_{\mathbf{e}} \sum_{\mathbf{a}} p(\mathbf{e}, \mathbf{a}|\mathbf{f}) \tag{9.37}$$

- Cf. LOGON ranking:
  2. Best path through graph, vs.
  3. Best translation

MBR

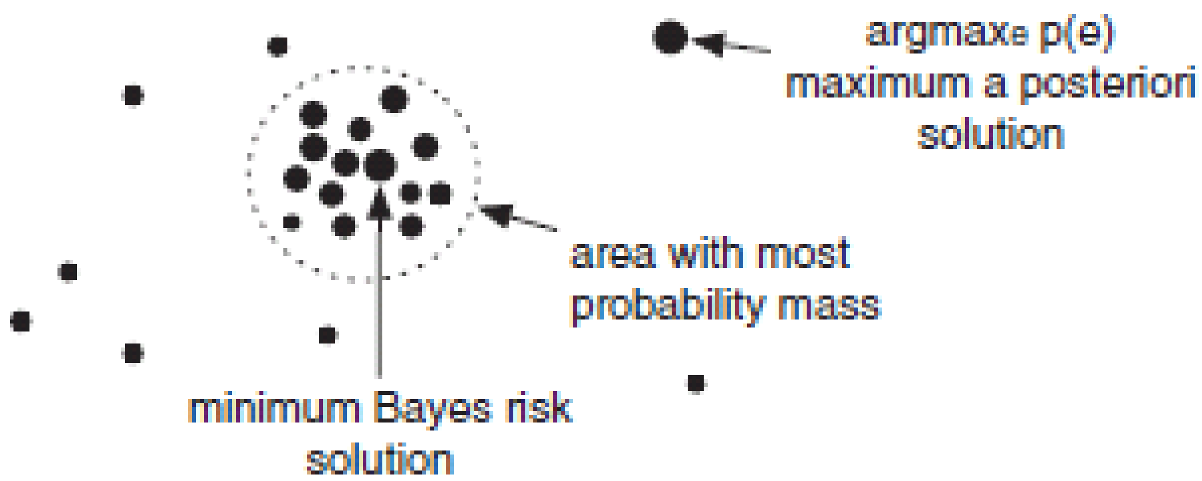


Figure 9.15 Minimum Bayes risk (MBR) decoding: This graph displays potential translations as circles, whose sizes indicate their translation probability. The traditional *maximum a priori* (MAP) decision rule picks the most probable translation. MBR decoding also considers neighboring translations, and favors translations in areas with many highly probable translations.

MBR

- Take into consideration distance to other (good) candidates
- How to measure distance:
  - BLEU?
  - Ideally, synonyms should come close together