Machine Translation for Multilingual Information Processing

Marcello Federico, Human Language Technology Research, Fondazione Bruno Kessler, Trento, Italy

TrebleCLEF Summer School on
Multilingual Information Access
Hotel Santa Croce in Fossabanda, Pisa, Italy
15-19 June 2009

(Statistical) Machine Translation for Multilingual Information Processing

Marcello Federico
FBK-irst Trento, Italy

Pisa, 15-19 June 2009

Outline

- Introduction
- Approaches
- Statistical MT
- Modeling
- Search
- Evaluation
- State-of-the-art
- Examples

Machine Translation

Wikipedia

Machine translation, often referred to by the acronym MT, is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another.

Preferred Definition

MT investigates the translation of "standard" language that can be systematically observed in ordinary communication – e.g. conversations, news, speeches, business letters, user manuals, etc. –. MT as a discipline is not interested in the translation of literature genres that express creative and sophisticated use of language. For several reasons, such kind of language is simply out of the scope of MT[1].

[1] A very interesting survey about the translation of literature work see U. Eco, "Experiences in Translation", 2000.

Introduction to MT

Why is Machine Translation so Difficult?

High quality human translation implies:

- deep and rich understanding of source language and text
- sophisticated and creative command of target language

Nowadays, feasible goals for machine translation are only tasks:

- for which a rough translation is adequate (gist translation)
- where a human post-editor can improve MT output (CAT)
- focusing on small linguistic domains (translators on PDAs)

In general, difficulty of translating depends on how similar the target and source languages are in their vocabulary, grammar, and conceptual structure.

Differences and Similarities of Languages

- **Universal communicative role** of language
  - names for people, words for talking about women, men, children
  - every language seems to have nouns and verbs
- **Differences/similarities across large classes of languages** [2]:
  - Morphological: one vs. many morphemes per words, agglutination vs. fusion
  - Syntactical: Subj-Verb-Obj structure (E) vs. SOV (J) vs. VSO (Irish)
  - Semantical: direction/manner of motion indicated by verb/satellites

  `the bottle floated out` (E) → `la botella salió flotando` (S)
- **Differences in specificity**, often peculiar to single languages:
  - Lexical: `informatique` (F) → `computer science` (E)
  - Syntactical: `she likes to sing` (E,v) → `sie singt gerne` (D,adv)
  - Semantical: `wall` (E) → `Wand/Mauer` (G, inside/outside)
- **Cultural Differences**: philosophical argument=is translation possible at all?

[2] A gentle introduction to MT is in D. Jurafsky and J. H. Martin, Speech and Language Processing, 2009.

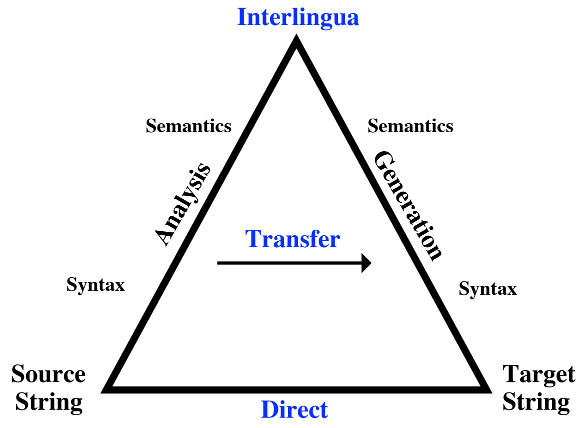
Approaches to MT

According to employed **linguistic representations**:

- **Direct model**: translate and re-order single words or n-grams
  - basically, no linguistic representation is used
- **Transfer model**: use explicit knowledge about language differences
  - analyze lexical and syntactic structure of source sentence
  - transfer structures from source to target language
  - generate corresponding sentence in the target language
- **Interlingua model**: extract the meaning and express it in the target language
  - analyze lexical, syntactical and semantical structure of source sentence
  - interpret the meaning into a canonical interlingua
  - generate the target sentence from the interlingua

Notice: required knowledge for the interlingua approach grows linearly with number of languages, rather than to the square.

Vauquois's Triangle



Approaches to MT

According to the **acquisition of models and knowledge**:

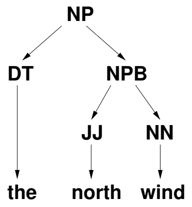
- **Hand-crafted**: knowledge for analysis, transfer, generation, meaning representation, or direct translation is manually developed
  - most of commercial MT systems fall in this category
  - requires lots of human labor and expertise
  - includes: rule-based MT
- **Machine-learned**: representations are implemented by mathematical models learnable from data, e.g. parallel corpora of human translations
  - much less human effort is needed
  - requires huge amounts of data, the more, the better!
  - includes: statistical MT and example-based MT

The two classifications are orthogonal

Transfer-Based MT

context-free grammar

NP	⟶	DT NPB
NPB	⟶	JJ NN
NPB	⟶	NN
	...	
DT	⟶	the
JJ	⟶	north
NN	⟶	wind
	...	



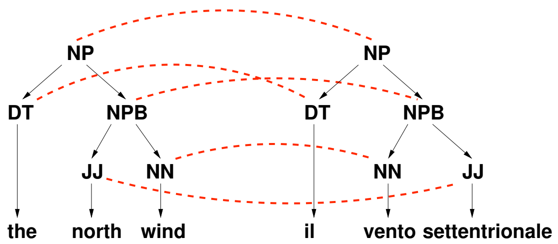
Transfer-Based MT

context-free grammar

NP	⟶	DT NPB
NPB	⟶	JJ NN
NPB	⟶	NN
	...	
DT	⟶	the
JJ	⟶	north
NN	⟶	wind
	...	

synchronous context-free grammar

NP	⟶	DT$_1$ NPB$_2$	/	DT$_1$ NPB$_2$
NPB	⟶	JJ$_1$ NN$_2$	/	NN$_2$ JJ$_1$
NPB	⟶	NN	/	NN
	...			
DT	⟶	the	/	il
JJ	⟶	north	/	settentrionale
NN	⟶	wind	/	vento
	...			



Statistical Machine Translation

- parallel texts

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

Statistical Machine Translation

- parallel texts

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

Statistical Machine Translation

- parallel texts

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- automatic word alignments

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

Statistical Machine Translation

- parallel texts

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- automatic word alignments

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- word translation probabilities

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

Statistical Machine Translation

- parallel texts

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- automatic word alignments

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

Statistical Machine Translation

- parallel texts

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- automatic word alignments

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- word translation probabilities

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

Statistical Machine Translation

- parallel texts

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- automatic word alignments

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

- word translation probabilities and target text probabilities

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

bigrams with **eastern**	counts	probs
eastern cool	5	0.05
eastern chilly	10	0.10
eastern wind	12	0.12
eastern breeze	7	0.07
eastern ...	...	...

Statistical Machine Translation

- given word translation probabilities and target text probabilities:

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

bigrams with **eastern**	counts	probs
eastern cool	5	0.05
eastern chilly	10	0.10
eastern wind	12	0.12
eastern breeze	7	0.07
eastern ...	...	...

- search over possible translations of the source sentence

un freddo vento da est

Statistical Machine Translation

- given word translation probabilities and target text probabilities:

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

bigrams with **eastern**	counts	probs
eastern cool	5	0.05
eastern chilly	10	0.10
eastern wind	12	0.12
eastern breeze	7	0.07
eastern ...	...	...

- search over possible translations of the source sentence

un freddo vento da est a cool eastern breeze

Statistical Machine Translation

- given word translation probabilities and target text probabilities:

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

bigrams with **eastern**	counts	probs
eastern cool	5	0.05
eastern chilly	10	0.10
eastern wind	12	0.12
eastern breeze	7	0.07
eastern ...	...	...

- search over possible translations of the source sentence

un freddo vento da est a cool eastern breeze
an eastern chilly wind

Statistical Machine Translation

- given word translation probabilities and target text probabilities:

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

bigrams with **eastern**	counts	probs
eastern cool	5	0.05
eastern chilly	10	0.10
eastern wind	12	0.12
eastern breeze	7	0.07
eastern ...	...	...

- search over possible translations of the source text

un freddo vento da est	a cool eastern breeze
	an eastern chilly wind
	a eastern cool wind
	a cold eastern wind
	an eastern chilly breeze
	...

Statistical Machine Translation

- given word translation probabilities and target text probabilities:

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

bigrams with **eastern**	counts	probs
eastern cool	5	0.05
eastern chilly	10	0.10
eastern wind	12	0.12
eastern breeze	7	0.07
eastern ...	...	...

- search over possible translations of the source text
  compute their probabilities or scores

un freddo vento da est	a cool eastern breeze	0.08
	an eastern chilly wind	0.10
	a eastern cool wind	0.09
	a cold eastern wind	0.12
	an eastern chilly breeze	0.05
	...	...

Statistical Machine Translation

- given word translation probabilities and target text probabilities:

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

bigrams with **eastern**	counts	probs
eastern cool	5	0.05
eastern chilly	10	0.10
eastern wind	12	0.12
eastern breeze	7	0.07
eastern ...	...	...

- search over possible translations of the source sentence
  compute their probabilities or scores and pick the most likely one

un freddo vento da est	a cool eastern breeze	0.08
	an eastern chilly wind	0.10
	a eastern cool wind	0.09
	a cold eastern wind	0.12
	an eastern chilly breeze	0.05
	...	...

Classical SMT Framework

Let $\mathbf{f}$ be any text in the source (foreign) language. The most probable translation $\hat{\mathbf{e}}$ is searched among texts in the target (English) language through the following statistical decision criterion[3]:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{f} \mid \mathbf{e}) \Pr(\mathbf{e}) \tag{1}$$

The **computational problems of SMT**:

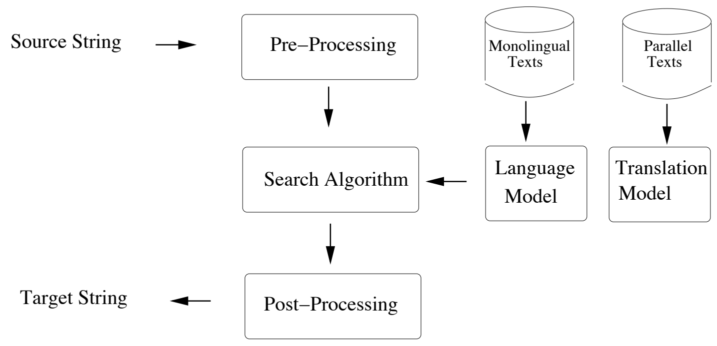
- language model: estimate probabilities $\Pr(\mathbf{e})$
- translation model: estimate probabilities $\Pr(\mathbf{f} \mid \mathbf{e})$
- search problem: carry out the optimization criterion (1)

Remarks

- all translation pairs are plausible, in principle, but have different probs
- although theory is presented with target English it is general

[3]Fundamental paper on SMT: P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, 1993.

Classical SMT Architecture



Search Criterion with Alignments

$$\begin{aligned} \texttt{vento freddo da est} &= \mathbf{f} \\ \texttt{eastern cool breeze} &= \arg\max_{\mathbf{e}} \Pr(\mathbf{f} \mid \mathbf{e}) \Pr(\mathbf{e}) \\ &= \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \Pr(\mathbf{e}) \end{aligned}$$

- Alignments are the (hidden) link between words of $\mathbf{f}$ and words of $\mathbf{e}$
- $\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$ is called Alignment Model
- Alignments permit to decompose or factorize the alignment model
- Notice that alignments map positions of $\mathbf{f}$ to positions of $\mathbf{e}$, not words!

$$\begin{array}{cccccccc} \texttt{vento}_1 & \texttt{freddo}_2 & \texttt{da}_3 & \texttt{est}_4 & \rightarrow & \texttt{null}_0 & \texttt{eastern}_1 & \texttt{cool}_2 & \texttt{breeze}_3 \\ 1 & 2 & 3 & 4 & \rightarrow & 0 & 1 & 2 & 3 \end{array}$$

Approximate Search Criterion

$$\begin{aligned} \hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \Pr(\mathbf{e}) \\ &\approx \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \ \Pr(\mathbf{e}) \end{aligned}$$

- instead of searching for the best translation
- compute the best translation and best alignment
- avoid summing over all alignments
- is this approximation reasonable?
  - yes, if for $\mathbf{f}$ and $\mathbf{e}$ there is only one probable alignment
  - not so good if the probable alignments are several
- a better approximation is summing over few most probable alignments

Model Factorization

Language and alignment models are factorized with two types of assumptions:

- bag-of-word model: $\Pr(a, b, c \mid e) \approx \Pr(a \mid e) \times \Pr(b \mid e) \times \Pr(c \mid e)$
- Markov chain model: $\Pr(a, b, c \mid e) \approx \Pr(a \mid e) \times \Pr(b \mid a, e) \times \Pr(c \mid b, e)$
- so probabilities are defined over smaller events and are easier to estimate

- **Alignment Model**: reordering model $\times$ lexicon model
  - $\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \Pr(\mathbf{a} \mid \mathbf{e}) \times \Pr(\mathbf{f} \mid \mathbf{a}, \mathbf{e})$
  - these models are further factorized and simplified!
  - there is a hierarchy of alignment models: Model 1 to Model 5[4]
  - we will review Model 1: the simplest alignment model
- **Language Model**: product of n-grams probabilities (e.g. 3-grams)
  - $\Pr(\mathbf{e}) = \Pr(e_1) \times \Pr(e_2 \mid e_1) ... \Pr(e_l \mid e_{l-2}, e_{l-1})$
  - the above trigram probabilities are smoothed to not over fit the training data

[4] See P. Brown et al., 1993.

The Current Approach to SMT

The so-called discriminative or **Log-Linear Model**[5] approach is based on:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \approx \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \qquad (2)$$

The posterior probability $\Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f})$ is determined through real valued feature functions $h_k(\mathbf{e}, \mathbf{f}, \mathbf{a})$, $k = 1 \dots M$, and takes the parametric form:

$$p_{\boldsymbol{\lambda}}(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \propto \exp\{\sum_k \lambda_k h_k(\mathbf{e}, \mathbf{f}, \mathbf{a})\} \qquad (3)$$

Important: this model includes the classical formulation if $M = 2$:

$$h_1(\mathbf{e}, \mathbf{f}, \mathbf{a}) = \log \Pr(\mathbf{e}),\ h_2(\mathbf{e}, \mathbf{f}, \mathbf{a}) = \log \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}),\ \lambda_1 = \lambda_2 = 1$$

[5] F. Och and H. Ney, Discriminative training and maximum entropy models for statistical machine translation, Proc. of ACL, 2002.

A Simple Alignment Model: Model 1

$$Pr(\mathbf{f} = f_1, \dots, f_m, \mathbf{a} = a_1, \dots, a_m \mid \mathbf{e} = \mathbf{e}_1, \dots, e_l)$$

is factorized in the following random processes:

1. Choose length m according to $p(m \mid l)$
2. Choose each a_j independently and at at random $\propto (l+1)^{-1}$
3. Choose each word f_j independently with $p(f_j \mid e_{a_j})$

Model 1 is poor to guide a MT search algorithm, but:

- It is robust to score translation hypotheses (semantic similarity?)
- The corresponding Translation Model is very efficient to compute
- The lexicon model $p(f \mid e)$ can be efficiently estimated from a parallel corpus

Models 1-5 can be trained with the open source code **GIZA++**

Model 1: Estimation

Let us assume that we have a parallel corpus with alignments:

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

Maximum Likelihood Estimation for discrete distributions:

$$\Pr(a \mid b) \approx \frac{\text{count}(a, b)}{\sum_a \text{count}(a, b)} = \frac{\text{count}(a, b)}{\text{count}(b)}$$

We can estimate translation probabilities by counting aligned word-pairs, e.g.:

$$\Pr(\text{chilly} \mid \text{freddo}) = \frac{\text{count}(\text{chilly}, \text{freddo})}{\text{count}(\text{freddo})} = \frac{1}{2} = 0.5$$

You have to imagine to use a very large parallel corpus indeed!

IBM Model 1: Viterbi Alignment

Let us assume that we probabilities $p(f \mid e)$ for all word-pairs, e.g.:

translations of **freddo**	counts	probs
chill	15	0.15
chilly	10	0.10
cold	43	0.43
cool	28	0.28
...	...	...

translations of **vento**	counts	probs
wind	59	0.59
breeze	26	0.26
...	...	...

Given a parallel corpus without alignments:

dalla serata di domani soffierà un freddo vento orientale
since tomorrow evening an eastern chilly wind will blow

un vento freddo da est interessa le Alpi
an eastern cool breeze affects the Alps

we can compute the most probable or **Viterbi alignment** of each sentence pair:

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} \Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e}) \text{ which gives:}$$

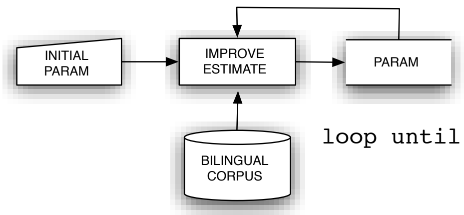
$$a_j^* = \arg\max_{i=0,1,\dots,l} p(f_j \mid e_i)$$

Estimation of IBM Models

How to train alignment models?

- we could use parallel data with alignments to train the model (MLE)
- we could compute the alignments through the model (Viterbi alignment)

Idea to solve this chicken & egg problem:



`loop until convergencence`

The above algorithms is called Expectation Maximization

EM Training Algorithm of Model 1

```
EM-Model2(F,m,E,l,S)
Init-Params(P); // Probabilities P[f,e]=p(f/e)
do
Reset-Counts(c); // Statistics c[f,e]
for s := 1 to S; // compute expected statistics
      do Expected-Counts(F[s],m[s],E[s],l[s],P,c);
for e ∈ ℰ;
      do tot[e]:=0;
         for f ∈ ℱ; // compute normalization
               do tot[e] := tot[e] + c[f,e];
         for f ∈ ℱ; // update parameters
               do P[f,e] := c[f,e]/tot[e];
until convergence
```


EM Training Algorithm of Model 1

```
Expected-Counts(F,m,E,l,P,c)
// Update statistics c[f,e] using P[]
for j := 1 to m;
      do t := 0;
         for i := 0 to l;
               do f=F[j]; e=E[i];
                  t := t + P[f,e];
         for i := 0 to l;
               do f:=F[j]; e:=E[i];
                  c[f,e] := c[f,e] + P[f,e] / t;
```


Trigram Language Model

The Language Model (LM)[6] gives the probability of $\mathbf{e} = e_1, e_2, \ldots, e_l$.

- The LM probability must also guess the length of $\mathbf{e}$
- We might assume an n-th order Markov chain factorization:

$$\Pr(\mathbf{e}_1^l) \approx \Pr(l) \cdot \Pr(e_1) \cdot \Pr(e_2 \mid e_1) \prod_{i=3}^{l} \Pr(e_i \mid e_{i-2},\ e_{i-1})$$

The above LM is called Trigram language model

- probabilities are estimated by smoothing relative frequencies of trigrams
- trigrams are collected on a huge text corpus in the target language
- Notice: LM probability can be computed incrementally on the target string
- LMs can be trained with **SRILM Toolkit** or **IRSTLM Toolkit**

[6] A gentle introduction to LMs is in D. Jurafsky and J. H. Martin, Speech and Language Processing, 2009.

ARPA File Format

N-grams probs (in the log space) can be computed by composing probs and back-off weights (≈ the prob of a new word after a given context):

```
\data
\ngram 1= 86700
\ngram 2= 1948935
\ngram 3= 2070512
\1-grams:
-2.88382        !       -2.38764
-2.94351        world        -0.514311
-6.09691        pisa        -0.15553
...
\2-grams:
-3.91009     world !     -0.351469
-3.91257     hello world  -0.24
-3.87582     hello pisa    -0.0312
..
\3-grams:
-0.00108858      hello world  !
-0.000271867     , hi hello  !
...
\end
```

```
logPr(!| hello pisa) = -0.0312 + logPr(!| pisa)
logPr(!| pisa) = -0.15553 - 2.88382
```


Decoding in SMT

Given a statistical alignment model, a language model, and a source sentence, the **task of the search procedure** is to find the most likely translation:

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} p(\mathbf{e}) \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

Often, we use the **Viterbi or maximum approximation**:

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} p(\mathbf{e}) \, \underset{\mathbf{a}}{max} \, p(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

Complexity of decoding depends on word-reordering:

- no word-reordering: polynomial (Viterbi algorithm)
- only local word-reordering: high-polynomial
- arbitrary word-reordering: NP-hard

Decoding Complexity

Decoding with Alignment Model 1 and a 2-gram Language Model:

$$\mathbf{e}^* = \operatorname*{argmax}_{l, e_1, e_2, \ldots, e_l} \underbrace{p(e_1 \mid \$) \cdot p(\$ \mid e_l) \cdot \prod_{i=2}^{l} p(e_i \mid e_{i-1})}_{\Pr(\mathbf{e})} \cdot \underbrace{\frac{p(m \mid l)}{(l+1)^m} \cdot \prod_{j=1}^{m} \sum_{i=1}^{l} p(f_j \mid e_i)}_{\sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}$$

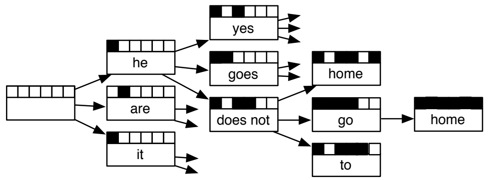
- Search might be limited to a fixed range of lengths for $\mathbf{e}$, e.g. $l \leq 2m$
- Computing probabilities of AM and LM is fast
- If any f has at most k translations, search space size is $O(k^{2m})$
- Decoding with M1 is NP-hard, i.e. almost no hope for an efficient algorithm[7].
  - The proof uses a reduction from the Hamiltonian Circuit Problem
- Approximations: beam-search algorithm + limited word-reordering

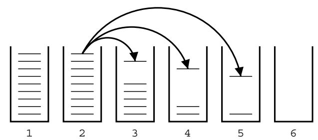
[7] Result presented in K. Knight, "Decoding Complexity in Word-Replacement Translation Models", Computational Linguistics, 1999.

Multi-stack Search with Dynamic Programming

Basic steps of search:

1. pick hyp. from one stack
2. cover new source positions
3. generate translation options
4. score hypotheses
5. recombine hypotheses
6. put them into stacks
7. prune stacks





Stack N contains hypotheses covering N positions

Search is performed over phrases rather than single words, ...

Solutions for efficiency

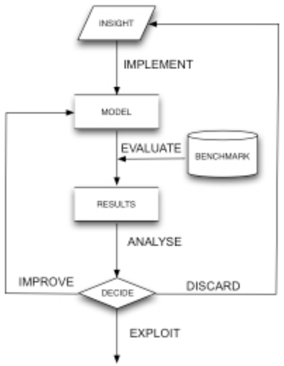
- **Constraints** to reduce expanded theories:
  - reordering constraints: limit number of allowed permutations
  - lexicon pruning: keep just most probable word translations
- **Beam search** to prune out less promising partial theories:
  - threshold pruning: keep just theories close to the best theory in the stacks
  - histogram pruning: keep at most M theories in the stacks
- **Memory optimization**:
  - garbage partial theories without successors
- **Efficient representation and use of**
  - Language Model probabilities: pruning, quantization, caching
  - Phrase-translation tables: pruning, quantization

Search with phrase-based translation can be performed with the **Moses Toolkit**[8].

[8] P. Koehn, at al., "Moses: Open Source Toolkit for Statistical Machine Translation", Proc. ACL 2007.

The Importance of Performance Evaluation

Experimental research in HLT is conducted according to the following cycle



Evaluation bottleneck: MT developers need to monitor the effect of daily changes to their systems in order to weed out bad ideas from good ideas!

Evaluating MT Performance

How do we evaluate the output of a MT system?

- **Human MT evaluation**:
  - criteria: adequacy, fidelity, and fluency
  - pros: very accurate, high quality
  - cons: expensive and slow
- **Automatic MT evaluation**:
  - criteria: similarity with respect to one or more human translations
  - pros: cheap, quick, correlates with human judgments
  - cons: correlation is not always high, scores are not comparable across tasks

  Example
  - BLEU: compute weighted sum of counts of the matching n-grams between output and references

The State of the Art

- **SMT is now a very competitive technology**
  - in many evaluations SMT outperformed rule-based MT
  - commercial systems perform likely better when not enough data are available
- Interest in SMT revamped around **seminal work at IBM in early 90'**
  - indeed the whole thing was started by **Warren Weaver in 1949**
- **Best performing SMT systems** use either:
  - brute force direct translation exploiting huge amounts of data
  - combination of direct translation and syntax-driven models
- **Automatic evaluation** metrics have dramatically boosted research in SMT:
  - model training directly optimizes the evaluation metric
- Several **evaluation campaigns** are organized every year:
  - NIST: news texts - Chi/Ara to Eng (2002-)
  - IWSLT: travelers speech - Chi/Jap/Ara/Ita to Eng (2004-)
  - MT Workshop: European parliament, news EU languages (2005-)

Example 1: Arabic English

Human	Dubai 2 - 7 (AFP) - The Secretary-General of the United Nations Kofi Annan said he would donate the international Zayed Prize for the Environment , which he received on Monday night in Dubai worth 500000 dollars , to setup a foundation for agriculture and educating girls in Africa .
Machine	Dubai 2-7 (AFP) - United Nations Secretary-General Kofi Annan said that the award will Zayed International Environment, which received Monday evening in Dubai worth 500,000 dollars to establish an institution for agriculture and education of girls in the African continent.

Example 2: Chinese English

Human	Today was the Catholic Church's annual " Life Day " . Pope Benedict XVI delivered a speech in St . Peter's Basilica , in which he criticized that the hedonism of wealthy society impairs the Christian value system of respect for life , and he strongly condemned abortion and euthanasia .
Machine	Today is the "life" of the Catholic Church once a year, when 16 of the pope delivered a speech in St. Peter's cathedral, criticized the joy of an affluent society, undermine the values of the Christian faith to respect life, and strongly condemned euthanasia and abortion.

Example 1: Arabic English

Human	Dubai 2 - 7 (AFP) - The Secretary-General of the United Nations Kofi Annan said he would donate the international Zayed Prize for the Environment , which he received on Monday night in Dubai worth 500000 dollars , to setup a foundation for agriculture and educating girls in Africa .
Machine	Dubai 2-7 (AFP) - United Nations Secretary-General Kofi Annan said that the award ~~will~~ Zayed International Environment, which ... he ... received ... on... Monday evening in Dubai worth 500,000 dollars ... , will be donated ... to establish an institution for agriculture and education of girls in the African continent.

Example 2: Chinese English

Human (?)	Today was the Catholic Church's annual " Life Day " . Pope Benedict XVI delivered a speech in St . Peter's Basilica , in which he criticized ~~that~~ the hedonism of ...our... wealthy society ...which... impairs the Christian value system of respect for life , and he strongly condemned abortion and euthanasia .
Machine	Today is the "life ..day..." of the Catholic Church ~~once a year~~, ~~when 16 of~~ the pope delivered a speech in St. Peter's cathedral, ...he... criticized the joy of an affluent society, ... that... undermines the values of the Christian faith to respect life, and strongly condemned euthanasia and abortion.

Thank you